Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

Anton Schiela

State constrained optimal control problems with states of low regularity [1]

[1] Supported by the DFG Research Center Matheon "Mathematics for key technologies"

ZIB-Report 08-24 (June 2008)

State constrained optimal control problems with states of low regularity [†]

Anton Schiela

June 17, 2008

Abstract

We consider first order optimality conditions for state constrained optimal control problems. In particular we study the case where the state equation has not enough regularity to admit existence of a Slater point in function space. We overcome this difficulty by a special transformation. Under a density condition we show existence of Lagrange multipliers, which have a representation via measures and additional regularity properties.

AMS MSC 2000: 49K20

Keywords: optimal control, state constraints, optimality conditions

1 Introduction

First order optimality conditions for optimal control problems subject to partial differential equations have been studied for a long time, successfully for large classes of problems. For an excellent overview we refer to the text book [19], for original papers in the pointwise state constrained elliptic case cf. [5, 6, 1].

The main structural assumption in the analysis of state constrained problems is the existence of a *Slater point*, which lies in the interior of the feasible set. Usual approaches require this to be satisfied with respect to the *norm topology* of the space of states. In the case of pointwise state constraints this holds only if $\|\cdot\|_\infty$ or a stronger norm is used. For cases were the coercivity of the functional and the properties of the PDE are strong enough to guarantee bounded states, optimality systems were derived and precise conclusions on the structure and the regularity of the dual variables were drawn. For the remaining cases not much is known. The only available result in this direction does not exploit the structure of a Slater point at all, which leads to poor conclusions. Similar issues arise in the analysis of bounds on the gradient of the state.

The purpose of this work is to close the remaining theoretical gap between these two extreme cases. Our idea is to introduce two separate topological frameworks.

[†] Supported by the DFG Research Center MATHEON "Mathematics for key technologies"

A *full* topological framework is needed for existence of a minimizer and a *restricted* framework is needed for a Slater condition. Then we transform the problem onto a space, where a Slater condition is satisfied *by construction*. By these techniques we can derive first order optimality conditions and similar regularity results as in [5, 6, 1] under weaker topological assumptions. In particular, if the *restricted* control space is dense in the *full* control space, the Lagrange multipliers of the state constraints still correspond to regular measures. In addition, these Lagrange multipliers are regular enough to be applied to all feasible, possibly discontinuous or unbounded states.

For simplicity we concentrate on the convex setting, remarking that smooth non-convex problems can usually be reduced locally to convex problems by linearization. For the sake of wide applicability and to clarify the analytic structure of the problem our analysis is performed in an abstract framework. We illustrate the application of our theory with an example from boundary control.

Acknowledgment The author wants to thank Prof. Dr. Fredi Tröltzsch for discussions on the subject, which gave rise to a couple of new ideas and motivated this thorough revision and extension of the preprint [18].

2 An Abstract Optimal Control Problem

Consider a convex optimal control problem of the following form:

$$\min_{(u,y)\in U\times Y} j(u,y) \quad s.t. \quad Ay - Bu = 0, \quad y \in \mathcal{Y}, \quad u \in \mathcal{U}. \tag{P}$$

As indicated, we will use two analytic frameworks for our problem.

Framework 2.1. (Full Framework) Let U, Y be normed spaces and R be a linear space. $Z := U \times Y$.

(i) Let $A : Y \supset \operatorname{dom} A \to R$ be an injective linear operator.

(ii) Let $B : U \to R$ be a linear operator.

(iii) There is a continuous "control-to-state“ mapping $S : U \to \operatorname{dom} A \subset Y$ such that $AS = B$.

(iv) Let $j : Z \to \mathbb{R}$, let $\mathcal{U} \subset U$ and $\mathcal{Y} \subset Y$. Define the feasible subspace

$$\mathcal{K} := \{(u,y) \in U \times \operatorname{dom} A : Ay - Bu = 0\} = \{(u,y) \in U \times \operatorname{dom} A : y = Su\}$$

and the feasible subset $\mathcal{Z} := (\mathcal{U} \times \mathcal{Y}) \cap \mathcal{K}$.

Observe that A, which models a *differential operator*, does not have to be defined on all of Y, but has its own domain of definition $\operatorname{dom} A$. This gives us additional flexibility to choose Y, as we will demonstrate in Section 5.

Framework 2.2. (Restricted Framework) Let Y_∞ and R_∞ be normed spaces. Let $U_\infty \subset U$ be a linear subspace of U. $Z_\infty := U_\infty \times Y_\infty$.

(i) There exists a continuous injective embedding $Y_\infty \hookrightarrow Y$, and an injective embedding $R_\infty \hookrightarrow R$. Via the ranges of these embeddings we will identify Y_∞ and R_∞ with linear subspaces of Y and R, respectively.

(ii) Let $A_\infty : Y_\infty \supset \operatorname{dom} A_\infty \to R_\infty$ be a linear operator with $\operatorname{dom} A_\infty \subset \operatorname{dom} A$ and $A_\infty = A$ on $\operatorname{dom} A_\infty$.

(iii) S maps U_∞ into $\operatorname{dom} A_\infty$, i.e., $S\,U_\infty \subset \operatorname{dom} A_\infty \subset Y_\infty$.

Our analysis will take place within these two frameworks, and the assumptions and notations introduced there will be in force throughout the whole paper. All other assumptions will be referenced explicitely, when needed.

To be able to show existence of minimizers we will need additional assumptions that hold within Framework 2.1. In particular, coercivity of j is crucial for existence.

Assumption 2.3. (Reflexivity, closedness, convexity, coercivity)

(i) U and Y are reflexive.

(ii) $\mathcal{U}$ and $\mathcal{Y}$ are closed, convex and non-empty. The feasible set $\mathcal{Z}$ is non-empty.

(iii) The functional j is convex and lower semi-continuous. It is coercive on $\mathcal{Z}$, i.e. for every sequence z_k in $\mathcal{Z}$, $\|z_k\|_Z \to \infty$ implies $j(z_k) \to \infty$.

Under Assumption 2.3 we will show later that our problem admits a minimizer, and it is actually possible to derive some sort of first order optimality conditions, however with poor conclusions. Most problems exhibit far more additional structure that we want to exploit in our analysis. For this we state the following additional regularity assumptions that hold only within the restricted Framework 2.2.

Assumption 2.4. (Restricted regularity)

(i) Y_∞ is complete.

(ii) There is $\tau > 0$ and $(\check{u}, \check{y}) \in \mathcal{Z} \cap Z_\infty$ such that $\check{y} + y \in \mathcal{Y}$ for all $y \in Y_\infty$ with $\|y\|_{Y_\infty} \leq \tau$ (Slater condition).

(iii) For every $u \in U$ there are $\lambda > 0$, $u_\infty \in U_\infty$, $u_{ad} \in \mathcal{U}$ with $u = \lambda(u_\infty + u_{ad})$.

In short, Framework 2.1 and Assumption 2.3 describe a setting, where the functional j is *coercive*, but a regularity condition fails to hold. The setting fixed in Framework 2.2 and Assumption 2.4 yields *regularity*, but no coercivity.

As we want to derive an abstract adjoint PDE, we need a regularity condition for the differential operator (cf. Appendix A).

Assumption 2.5. Y_∞ and R_∞ are complete and the operator A_∞ is closed, densely defined, and has closed range.

As an example, consider a pointwise state constrained linear quadratic optimal control problem (for precise example cf. Section 5). Then, to assert coercivity, U must be an L_2-space, and Y has to be chosen *sufficiently large*, to guarantee continuity of $S : U \to Y$. Then Framework 2.1 is fixed, and Assumption 2.3 holds. In contrast, Assumption 2.4*(ii)* can only be fulfilled, if Y_∞ is *sufficiently small.* In the state constrained case Y_∞ usually has to be a space of continuous functions. Then U_∞ and R_∞ have to be chosen accordingly.

In some cases, the regularity theory of the PDE shows that the choices $Y_\infty = Y$ and $U_\infty = U$ are possible. These cases are well analysed (cf. e.g. [5, 6, 1]) and precise structural results are known. Observe that Assumption 2.4*(iii)* is trivially fulfilled in this case.

The remaining cases include Neumann boundary control in three space dimensions, many parabolic control problems, and problems with bounds on the gradient of the state. In Section 5 we will consider three dimensional boundary control in detail. Analysis of these cases has been done (cf. e.g. [6, 1, 7, 8]), but only for special choices of j and $\mathcal{U}$ to obtain sufficiently strong coercivity properties.

This work explores the case, where both settings do *not* coincide with the aim to exploit as much structure are possible. Of particular interest is the case where U_∞ is *dense* in U. We will show that the conclusions of the case $(Z, R) = (Z_\infty, R_\infty)$ essentially extend to this more general case. Because U is often chosen as an L_p-space in applications, a suitable dense subspace is easily found.

Without a density assumption the conclusions become considerably weaker. The extreme case in this direction $(Z_\infty, R_\infty) = (\{0\}, \{0\})$ has been considered previously (cf. e.g. [12]). Assumption 2.4*(i)-(ii)* and 2.5 hold trivially in this case, while Assumption 2.4*(iii)* is only valid for $\mathcal{U} = U$. However, to get a true insight into the structure of a particular class of problems it is vital to choose the restricted framework and in particular U_∞ as large as possible. The larger U_∞, Y_∞ and R_∞, the stronger the results.

3 Main results

For the statement of our main results, in particular first order optimality conditions for the problem (P), we will use basic concepts of the theory of unbounded operators and convex analysis (cf. Appendix A and B).

3.1 Existence of minimizers

With the help of indicator functions (67) we can rewrite (P) as an unconstrained optimal control problem.

$$\min_{z=(u,y)\in Z} F(z) := j(z) + \iota_{\mathcal{Z}} = j(z) + \iota_{\mathcal{U}}(u) + \iota_{\mathcal{Y}}(y) + \iota_{\mathcal{K}}(z). \tag{1}$$

Now $F : Z \to \overline{\mathbb{R}}$ is an extended real valued function. Equivalence to (P) follows, because $\iota_{\mathcal{Z}}(z) = 0$, if z is feasible and $+\infty$ otherwise.

Theorem 3.1. *If Assumption 2.3 holds, then* (P) *admits a minimizer* $z_{opt} \in \mathcal{Z}$.

Proof. The operator $(-S, I) : U \times Y \to Y$ is continuous and thus $\mathcal{K} = \ker(-S, I)$ is closed. Hence, by Assumption 2.3*(ii)* $\mathcal{Z}$ is closed and convex as the intersection of the closed and convex sets $\mathcal{K}$ and $\mathcal{U} \times \mathcal{Y}$, and non-empty. Thus $\iota_{\mathcal{Z}}$ is convex, lower semi-continuous and proper. By Assumption 2.3*(iii)*, j is convex, lower semi-continuous and finite on Z and coercive on $\mathcal{Z}$.

So all in all, $F = j + \iota_{\mathcal{Z}}$ is convex, lower-semi continuous, proper, and coercive on the space Z, which is reflexive by Assumption 2.3*(i)*. Thus we can apply the main existence theorem of convex optimization (cf. e.g. [9, Proposition II.1.2]), which yields existence of a minimizer $z_{opt} \in Z$ of (P). □

3.2 A tailored function space

It turns out that the analysis of (P) and its first order optimality conditions requires a particularly tailored function space $\tilde{Y}$ for the states. It is constructed to contain all feasible solutions and to reflect the regularity structure of the problem in view of Assumption 2.4*(ii)*.

Definition 3.2. Define the following subspace of Y:

$$\tilde{Y} := \operatorname{ran} S + Y_\infty \subset Y, \tag{2}$$

and the following functional on $\tilde{Y}$:

$$\|y\|_{\tilde{Y}} := \inf\{\|u\|_U + \|w\|_{Y_\infty} : u \in U, w \in Y_\infty, y = Su + w\}. \tag{3}$$

The next proposition states the unsurprising result that $(\tilde{Y}, \|\cdot\|_{\tilde{Y}})$ is a normed space:

Proposition 3.3. *The functional* $\|\cdot\|_{\tilde{Y}}$ *defines a norm on* $\tilde{Y}$ *and there exist the injective continuous embeddings* $\tilde{E} : Y_\infty \hookrightarrow \tilde{Y}$ *and* $E : \tilde{Y} \hookrightarrow Y$. *The control-to-state mapping* S *is continuous as a mapping* $\tilde{S} : U \to \tilde{Y}$.

If additionally U_∞ *is dense in* U *then the embedding* $Y_\infty \hookrightarrow \tilde{Y}$ *is dense, and if* $\operatorname{dom} A_\infty$ *is dense in* Y_∞*, then the following operators are densely defined*

$$\begin{aligned} \tilde{A} : \tilde{Y} \supset \operatorname{dom} A_\infty &\to R_\infty \\ y &\mapsto A_\infty y, \end{aligned} \tag{4}$$

$$\begin{aligned} \tilde{B} : U \supset U_\infty &\to R_\infty \\ u &\mapsto Bu. \end{aligned} \tag{5}$$

Proof. First we show that (3) is a semi-norm. Clearly, $\|\cdot\|_{\tilde{Y}}$ is non-negative, positively homogenous, and $\|y\|_{\tilde{Y}} < \infty$ for all $y \in \tilde{Y}$.

To see that the triangle inequality holds, let $\varepsilon > 0$, and choose for $y_i \in \tilde{Y}$, $i = 1, 2$, $u_i \in U$ and $w_i \in Y_\infty$, such that $y_i = Su_i + w_i$ and $\|u_i\|_U + \|w_i\|_{Y_\infty} \leq \|y_i\|_{\tilde{Y}} + \varepsilon$.

Then

$$\begin{aligned}\|y_1+y_2\|_{\tilde Y} &\le \|u_1+u_2\|_U+\|w_1+w_2\|_{Y_\infty} \le \|u_1\|_U+\|u_2\|_U+\|w_1\|_{Y_\infty}+\|w_2\|_{Y_\infty}\\ &\le \|y_1\|_{\tilde Y}+\|y_2\|_{\tilde Y}+2\varepsilon.\end{aligned}$$

Because ε was arbitrary, the triangle inequality follows, and $\|\cdot\|_{\tilde Y}$ is a semi-norm.

Similarly, to show $\|\cdot\|_Y \le c\,\|\cdot\|_{\tilde Y}$ choose for $y\in\tilde Y$, $u\in U$ and $w\in Y_\infty$ such that $y=Su+w$ and $\|u\|_U+\|w\|_{Y_\infty}\le \|y\|_{\tilde Y}+\varepsilon$. Then Framework 2.1(*iii*) and 2.2 (*i*) yield

$$\|y\|_Y \le \|Su\|_Y+\|w\|_Y \le \|S\|\,\|u\|_U + c\,\|w\|_{Y_\infty} \le \max\{\|S\|, c\}(\|y\|_{\tilde Y}+\varepsilon).$$

Thus $E:\tilde Y\hookrightarrow Y$ is continuous. This implies also that, $\|\cdot\|_{\tilde Y}$ is not only a semi-norm, but a norm on $\tilde Y$ and E is injective: if $\|y\|_{\tilde Y}=0$, then also $\|y\|_Y=0$, and thus $y=0$, because $\tilde Y$ was defined as a subspace of Y.

Continuity of $\tilde S: U\to\tilde Y$ immediately follows from the definition of $\|\cdot\|_{\tilde Y}$ (setting $w=0$), which yields $\|\tilde Su\|_{\tilde Y}\le\|u\|_U$. Further, $\|y\|_{\tilde Y}\le\|y\|_{Y_\infty}$ follows from the choice $u=0$ in (3), and thus the embedding $\tilde E: Y_\infty\hookrightarrow\tilde Y$ is continuous. Injectivity of $\tilde E$ follows, because the embedding $Y_\infty\hookrightarrow Y$ is injective due to Framework 2.2*(i)* and $\tilde Y$ is a subspace of Y.

By Framework 2.2(*iii*) and (*ii*) $SU_\infty\subset\operatorname{dom}A_\infty$ and $A=A_\infty$ on $\operatorname{dom}A_\infty$. We compute for $u\in U_\infty$: $ASu=A_\infty Su=Bu$, and thus, because A_∞ maps into R_∞, $Bu\in R_\infty$. Hence, $\tilde B$ is well defined in (5). If U_∞ is dense in U, then clearly $\tilde B$ is densely defined. For the remaining density assertions let $y\in\tilde Y$. Then there is $u\in U$ and $w\in Y_\infty$ with $y=Su+w$. Further, there is a sequence u_k in U_∞ with $\|u_k-u\|_U\to 0$. Consequently $y_k:=Su_k+w\in Y_\infty$, and $\|y-y_k\|_{\tilde Y}\to 0$ by (3). Hence, Y_∞ is dense in $\tilde Y$, which implies also that $\tilde A$ is densely defined. □

Finally, we define for later reference the following space and a continuous embedding

$$\tilde Z := U\times\tilde Y, \qquad E_Z := (I,E):\tilde Z\hookrightarrow Z. \tag{6}$$

Note that $S=E\tilde S$ and that both operators are algebraically equivalent and thus might as well be identified. However, as we will consider their adjoints later, a notational distinction seems to be appropriate for the sake of clarity. Similarly, we write the embeddings E_Z, $\tilde E$ and E explicitely, because we will use their adjoints. Adjoints of embeddings usually have an interpretation as restrictions of linear functionals to subspaces.

Further, we define $\tilde{\mathcal Y}:=\mathcal Y\cap\tilde Y$. More formally, $\tilde{\mathcal Y}$ is the pre-image of $\mathcal Y$ with respect to E. Because E is continuous, $\tilde{\mathcal Y}$ is closed in $\tilde Y$. Observe that the subspace $\mathcal K$ defined in framework 2.1*(iv)* is contained in $\tilde Z$. Thus $\tilde Z$ contains the feasible set of (P) and thus also its minimizers.

3.3 First order optimality conditions

Now we can formulate our main result, which is the most important special case of Theorem 4.8, proved in Section 4. Observe that all quantities are well defined by Proposition 3.3, in particular $\tilde{A}^*$ and $\tilde{B}^*$ (cf. Appendix A.2).

Theorem 3.4. *Suppose that Assumptions 2.3-2.5 hold and that U_∞ is dense in U. Then $z_{opt} = (u_{opt}, y_{opt}) \in Z$ is a minimizer of* (P) *if and only if $z_{opt} \in \mathcal{Z}$ and the following system of equations has a solution $(j^*, u^*, y^*, p) \in Z^* \times U^* \times \tilde{Y}^* \times R_\infty^*$:*

$$E^* j_y^* + y^* + \tilde{A}^* p = 0 \quad \textit{in } \tilde{Y}^* \tag{7}$$

$$j_u^* + u^* - \tilde{B}^* p = 0 \quad \textit{in } U^* \tag{8}$$

$$\langle y^*, \tilde{y} - y_{opt} \rangle \leq 0 \quad \forall \tilde{y} \in \tilde{\mathcal{Y}} \tag{9}$$

$$\langle u^*, u - u_{opt} \rangle \leq 0 \quad \forall u \in \mathcal{U} \tag{10}$$

$$p \in \operatorname{dom} \tilde{B}^* \cap \operatorname{dom} \tilde{A}^* \subset R_\infty^*, \quad j^* = (j_u^*, j_y^*) \in \partial j(z_{opt}). \tag{11}$$

Proof. Compare (35)-(40) and (7)-(11). Obviously Theorem 3.4 follows from Theorem 4.8, if we can show that there is $p \in \operatorname{dom} \tilde{B}^* \cap \operatorname{dom} \tilde{A}^*$, such that $(-\tilde{S}^* a^*, a^*) = (-\tilde{B}^* p, \tilde{A}^* p)$.

Since U_∞ is dense in U and by Assumption 2.5, Proposition 3.3 yields that $\tilde{A}$ and $\tilde{B}$ are densely defined (thus $\tilde{A}^*$ and $\tilde{B}^*$ are well defined, cf. Appendix A.2), and Y_∞ is dense in $\tilde{Y}$. Equation (40) yields $\langle a^*, y \rangle = \langle r^*, A_\infty y \rangle = \langle r^*, \tilde{A} y \rangle$ for all $y \in Y_\infty$. Because $a^* \in \tilde{Y}^*$, also $\langle r^*, \tilde{A} \cdot \rangle$ is continuous on the dense subspace Y_∞ of $\tilde{Y}$. Hence, by definition of $\operatorname{dom} \tilde{A}^*$, $r^* \in \operatorname{dom} \tilde{A}^*$, and $\tilde{A}^* r^* = a^*$, because this equality holds on a dense subset of $\tilde{Y}$.

Moreover, Proposition 4.7 yields $\langle \tilde{S}^* a^*, u \rangle = \langle r^*, Bu \rangle = \langle r^*, \tilde{B} u \rangle$ for all $u \in U_\infty$, which is again dense in U. Similarly as above we obtain $r^* \in \operatorname{dom} \tilde{B}^*$ and $\tilde{S}^* a^* = \tilde{B}^* r^*$. Hence, we may set $p = r^*$. □

Theorem 3.4 is a structural result in the first place, which also implies regularity assertions for the quantities p and y^*. In Section 5 we will consider an example where the space R_∞^* is a Sobolev space $W^{1,q'}(\Omega)$. Hence, p is smooth (this explains, why the "adjoint state" p is perceived as a function, rather than as a functional). Usually one can conclude also smoothness of the optimal control u_{opt} from this. This regularity result is independent of the representation of $\tilde{Y}^*$, which we will discuss in the following section.

3.4 Representations of the space of Lagrange multipliers

Theorem 3.4 states existence of Lagrange multipliers in an abstract dual space $\tilde{Y}^*$. It is thus interesting to consider representations of $\tilde{Y}^*$. To keep the discussion concrete, we will consider the case $Y_\infty = C(Q)$ for a compact set Q, which implies by the Riesz representation theorem that $Y_\infty^* \cong M(Q)$, the space of Radon measures on Q.

Proposition 3.5. *Let $Y_\infty = C(Q)$ and suppose that the embedding $\tilde{E} : Y_\infty \hookrightarrow \tilde{Y}$ is dense. There is the continuous injective embedding*

$$\tilde{E}^* : \tilde{Y}^* \hookrightarrow M(Q).$$

In particular, there is the following representation for $\tilde{Y}^*$:

Let $y^* \in \tilde{Y}^*$ and $\mu = \tilde{E}^* y^*$. For each $y \in \tilde{Y}$ there are sequences y_k in $C(Q)$ with $y_k \to y$ in $\tilde{Y}$, and it holds independently of the choice of the sequence

$$\langle y^*, y\rangle = \lim_{k\to\infty} \int_Q y_k \, d\mu, \tag{12}$$

$$\left| \langle y^*, y\rangle - \int_Q y_k \, d\mu \right| \le \|y^*\|_{\tilde{Y}^*} \, \|y - y_k\|_{\tilde{Y}} \, . \tag{13}$$

Proof. By assumption there is a continuous embedding $\tilde{E} : Y_\infty \hookrightarrow \tilde{Y}$, which has dense range. Its (continuous) adjoint mapping $\tilde{E}^* : \tilde{Y}^* \hookrightarrow Y_\infty^* \cong M(Q)$ is thus injective by Theorem A.2.

Next, (13) follows from

$$\left| \langle y^*, y\rangle - \int_Q y_k \, d\mu \right| = |\langle y^*, y\rangle - \langle y^*, y_k\rangle| \le \|y^*\|_{\tilde{Y}^*} \, \|y - y_k\|_{\tilde{Y}} \, .$$

Finally, density of $Y_\infty \hookrightarrow \tilde{Y}$ and (13) imply (12). □

Remark 3.6. Density of $Y_\infty \hookrightarrow \tilde{Y}$ and thus injectivity of $\tilde{E}^*$ is crucial for the regularity of $\tilde{Y}$. A notorious example with missing injectivity is the "embedding" $L_\infty(Q)^* \hookrightarrow M(Q)$. Although its continuity implies that each element of L_∞^* acts as a measure if applied to a continuous function, the representation of L_∞^* (as space of countably additive set functions) is less regular than $M(Q)$. This is due to the very large and irregular kernel of this embedding: each measure corresponds to a large affine subspace of L_∞^*. In contrast, $\ker \tilde{E}^* = 0$, so each measure corresponds to at most one element of $\tilde{Y}^*$ and thus actually serves as a representation of this element.

Clearly, if $y \in \tilde{Y}$ is a continuous function, then $\langle y^*, y\rangle$ is represented via an integral. Proposition 3.5 states that for all other elements of $\tilde{Y}$ the evaluation of $\langle y^*, y\rangle$ can be done via converging sequences of integrals, but not *directly* via an integral in general.

Such a situation occurs surprisingly often at the interface between measure theory and functional analysis. The interested reader is referred to [3, 10] for an account on their interplay. Perhaps the most prominent example is the Fourier transformation, which is defined as an integral transform only on $L_1(\mathbb{R}^d)$, but isometric with respect to the L_2-norm. This allows to extend the corresponding linear operator to an isometry on $L_2(\mathbb{R}^d)$, called Fourier-Plancherel transformation, and use it there conveniently in a Hilbert space setting. Failure of the integral in the context of optimal control can be observed in [13]. Here a control problem is considered, where the optimal solution depends on the type of integral (Lebesgue or improper Riemann integral) used.

In some cases, however, we do have a direct representation via an integral. Here one has to take into account that the space Y is usually a space of equivalence classes of functions, and $\tilde{Y}$ is a subspace thereof. Thus, such a representation is only meaningful via a continuous "trace" operator $\gamma_\mu : \tilde{Y} \to L_1(\mu)$. Here is a simple case, where γ_μ exists.

Proposition 3.7. *Let $Y_\infty = C(Q)$ and suppose that $\tilde{E} : Y_\infty \hookrightarrow \tilde{Y}$ is dense. Assume that there is a constant C, such that*

$$\| |y| \|_{\tilde{Y}} \le C \, \|y\|_{\tilde{Y}} \quad \forall y \in C(Q). \tag{14}$$

Let $y^* \in \tilde{Y}^*$ and $\mu = \tilde{E}^* y^*$. If μ is positive, then there is a linear continuous trace operator

$$\gamma_\mu : \tilde{Y} \to L_1(\mu),$$

such that $\gamma_\mu(y) = y \; \forall y \in C(Q)$ and

$$\langle y^*, y \rangle = \int_Q \gamma_\mu(y) \, d\mu \quad \forall \, y \in \tilde{Y}. \tag{15}$$

Proof. For $y \in C(Q)$ define $\gamma_\mu(y) := y$. By (14) we have:

$$\|\gamma_\mu(y)\|_{L_1(\mu)} = \int_Q |\gamma_\mu(y)| \, d\mu \le \| |y| \|_{\tilde{Y}} \, \|y^*\|_{\tilde{Y}^*} \le C \, \|y\|_{\tilde{Y}} \, \|y^*\|_{\tilde{Y}^*} \, .$$

Hence, γ_μ is continuous on the dense subspace $C(Q) \subset \tilde{Y}$. Since additionally $L_1(\mu)$ is complete, this operator has a unique continuous extension (cf. e.g. [20, Satz II.1.5]) with the stated properties. □

We close this section by noting that there may be more sophisticated representation criteria, depending on the particular problem. Further, it may be an interesting task to analyse the subspace ran $\tilde{E}^*$ for a particular problem.

4 Proof of the main results

We will now proof first order optimality conditions for (P). We have to cope with the problem that the space Z is too large for a direct approach. Our strategy thus consists of three main steps. In Section 4.1 we transform (P) from Z to an equivalent problem on an auxiliary space X. In Section 4.2, which contains the core of the proof, we apply the sum-rule of convex analysis to the transformed version of (P). Characterization of the summands yields a dual equation in X^*. In Section 4.3 a back-transformation to $\tilde{Z}^*$ is performed.

In our proof we will use basic tools from the theory of unbounded operators and from convex analysis. For convenient reference we have gathered these tools in the appendix.

4.1 Transformation of the problem

We introduce the auxiliary space X and a transformation $\tilde{T} : X \to \tilde{Z}$. X and $\tilde{T}$ will be chosen, such that the feasible subspace $\mathcal{K}$ is transformed to the first component of X. This gives us the freedom to choose the second component of X according to our needs, namely as Y_∞ to exploit a Slater condition. We define the normed space X by

$$X := U \times Y_\infty, \quad \|\cdot\|_X = \|\cdot\|_U + \|\cdot\|_{Y_\infty}.$$

The components of X will *not* be interpreted as control and state. We will express this by the notational convention $x = (v, w)$. Rather, the connection between X and $\tilde{Z}$ is established by the following transformation ($\tilde{S}$ and $\tilde{E}$, are defined in Proposition 3.3).

Proposition 4.1. *There is the continuous transformation:*

$$\begin{aligned} &\tilde{T} : X \to \tilde{Z} \\ &(v, w) \mapsto (u, y) := (v, \tilde{S}v + \tilde{E}w). \end{aligned} \tag{16}$$

$\tilde{T}$ is injective and $\operatorname{ran} \tilde{T} \supset \mathcal{K} + Z_\infty$. *$\tilde{T}|_U : U \to \mathcal{K} \subset \tilde{Z}$ is an isomorphism.*

Proof. By construction of $\tilde{Y}$ (cf. Proposition 2), $\tilde{T}$ is well defined. If $\tilde{T}(v, w) = (0, 0)$, then $v = 0$ and $\tilde{S}v + \tilde{E}w = 0$, thus also $\tilde{E}w = 0$. Since $\tilde{E}$ is injective, $\tilde{T}$ is injective, too. Continuity of $\tilde{T}$ follows from

$$\|\tilde{T}(v, w)\|_{\tilde{Z}} = \|v\|_U + \|\tilde{S}v + \tilde{E}w\|_{\tilde{Y}} \le \|v\|_U + (\|v\|_U + c\,\|w\|_{Y_\infty}) \le c\,\|(v, w)\|_X .$$

Consider $\tilde{T}|_U$, which maps $(v, 0)$ to $(v, \tilde{S}v)$. As a restriction, $\tilde{T}|_U$ inherits injectivity and continuity from $\tilde{T}$. From Framework 2.1*(iv)* we obtain $\{(v, \tilde{S}v) : v \in U\} = \mathcal{K}$. Hence, $\operatorname{ran} \tilde{T}|_U = \mathcal{K}$. Thus, $\tilde{T}|_U : U \to \mathcal{K}$ is bijective and its inverse is continuous, because $\|\tilde{T}(v, 0)\|_{\tilde{Z}} = \|v\|_U + \|\tilde{S}v\|_{\tilde{Y}} \ge \|v\|_U = \|(v, 0)\|_X$.

If $z = (u, y) \in Z_\infty \subset \tilde{Z}$, then $\tilde{S}u \in Y_\infty$ by Framework 2.2*(iii)*. Hence, $y - Su \in Y_\infty$, thus $x := (u, y - Su) \in X$ and we compute

$$\tilde{T}x = (u, \tilde{S}u + \tilde{E}(y - Su)) = (u, \tilde{E}y) = (u, y) = z$$

and thus $\operatorname{ran} \tilde{T} \supset Z_\infty$. We have already shown that $\operatorname{ran} \tilde{T} \supset \mathcal{K}$, thus $\operatorname{ran} \tilde{T} \supset Z_\infty + \mathcal{K}$. □

We may also write our transformation in matrix form:

$$\tilde{T} = \begin{pmatrix} I & 0 \\ \tilde{S} & \tilde{E} \end{pmatrix} : U \times Y_\infty \to U \times \tilde{Y}.$$

The composition of $\tilde{T}$ with the continuous embedding $E_Z : \tilde{Z} \to Z$ defined in (6) yields the continuous transformation

$$T := E_Z \tilde{T} : X \to Z. \tag{17}$$

Proposition 4.2. *The following assertions are equivalent for $x \in X$:*

(i) Tx *is a minimizer of* $F : Z \to \overline{\mathbb{R}}$ *as defined in* (1).

(ii) x *is a minimizer of* $F \circ T : X \to \overline{\mathbb{R}}$.

(iii) $0 \in \partial(F \circ T)(x) \subset X^*$.

Proof. First of all *(ii)* $\Leftrightarrow$ *(iii)* by (68). To show *(i)* $\Leftrightarrow$ *(ii)*, we note that all minimizers of F are in $\mathcal{K}$, and T maps U onto $\mathcal{K}$ bijectively. Since $(F \circ T)(x) = F(Tx)$ equivalence of *(i)* and *(ii)* follows. □

For the back-transformation in Section 4.3 we need some results on $\tilde{T}^*$.

Proposition 4.3. *Consider the continuous adjoint operator of* $\tilde{T}$*:*

$$\begin{aligned} \tilde{T}^* &: \tilde{Z}^* \to X^* \\ (u^*, y^*) &\mapsto (v^*, w^*) := (u^* + \tilde{S}^* y^*, \tilde{E}^* y^*). \end{aligned} \tag{18}$$

(i) There are the following characterizations:

$$\ker \tilde{T}^* = \left\{ (-\tilde{S}^* y^*, y^*) \in \tilde{Z}^* : y^* \in \ker \tilde{E}^* \subset \tilde{Y}^* \right\} \tag{19}$$

$$\operatorname{ran} \tilde{T}^* = \left\{ (v^*, w^*) \in X^* : \langle w^*, w \rangle \le c \|\tilde{E} w\|_{\tilde{Y}} \ \forall w \in Y_\infty \right\}. \tag{20}$$

(ii) z^* *solves the equation* $\tilde{T}^* z^* = (v^*, w^*)$ *if and only if there is* $y^* \in \tilde{Y}$*, such that* $\tilde{E}^* y^* = w^*$ *and* $z^* = (v^* - \tilde{S}^* y^*, y^*)$.

(iii) If $\langle x^*, x \rangle = 0$ *for all* $x = (v, w)$ *satisfying* $\tilde{S} v + \tilde{E} w = 0$*, then there is* $y^* \in \tilde{Y}^*$*, such that* $z^* = (0, y^*)$ *and* $\tilde{T}^* z^* = x^*$.

Proof. Let $z^* = (u^*, y^*) \in \tilde{Z}^*$. Then (18) follows from the computation

$$\langle \tilde{T}^* z^*, x \rangle = \langle z^*, \tilde{T} x \rangle = \langle u^*, v \rangle + \langle y^*, \tilde{S} v + \tilde{E} w \rangle = \langle u^* + \tilde{S}^* y^*, v \rangle + \langle \tilde{E}^* y^*, w \rangle.$$

By (18) $(v^*, w^*) = 0$, iff $\tilde{E}^* y^* = 0$ and $u^* = -\tilde{S}^* y^*$. This yields (19).

For the characterization of $\operatorname{ran} \tilde{T}^*$ denote by M the set defined on the right hand side of (20). It follows from (18) that $\operatorname{ran} \tilde{T}^* \subset M$, because $w^* = \tilde{E}^* y^*$ and $y^* \in \tilde{Y}^*$.

Let in converse $x^* = (v^*, w^*) \in M$. By injectivity of $\tilde{E}$, w^* induces a continuous linear functional e^* on $\operatorname{ran} \tilde{E} \subset \tilde{Y}$ via the definition $\langle e^*, \tilde{E} w \rangle := \langle w^*, w \rangle$. By the Hahn-Banach theorem e^* can be extended continuously to a functional $y_0^* \in \tilde{Y}^*$, and it holds $w^* = \tilde{E}^* y_0^*$. Setting $z_0^* = (v^* - \tilde{S}^* y_0^*, y_0^*)$ we compute

$$\langle z_0^*, \tilde{T} x \rangle = \langle v^* - \tilde{S}^* y_0^*, v \rangle + \langle y_0^*, \tilde{S} v + \tilde{E} w \rangle = \langle v^*, v \rangle + \langle \tilde{E}^* y_0^*, w \rangle = \langle v^*, v \rangle + \langle w^*, w \rangle,$$

and hence $x^* = \tilde{T}^* z_0^* \in \operatorname{ran} \tilde{T}^*$, which implies $M \subset \operatorname{ran} \tilde{T}^*$ and thus (20). Linear algebra yields that the set of solutions of the equation $x^* = \tilde{T}^* z^*$ is $z_0^* + \ker \tilde{T}^*$. By (19) it is clear that $z_0^* + \ker \tilde{T}^*$ is the set of all z^* of the form $z^* = (v^* - \tilde{S}^* y^*, y^*)$, with $\tilde{E}^*y^* = \tilde{E}^*y_0^* = w^*$. This yields the characterization of the set of solutions *(ii)*.

Next, we show (iii). By definition of $\tilde{Y}$, every $\tilde{y} \in \tilde{Y}$ can be written in the form $\tilde{y} = \tilde{S}v + \tilde{E}w$. Let $\langle x^*, x\rangle = \langle v^*, v\rangle + \langle w^*, w\rangle = 0$ for all x that satisfy $\tilde{S}v + \tilde{E}w = 0$. Then for $\tilde{S}v + \tilde{E}w = \tilde{y} \in \tilde{Y}$ the expression $\langle v^*, v\rangle + \langle w^*, w\rangle$ depends only on $\tilde{y}$ and not on the choice of v and w. Thus we can define a linear functional y^* on $\tilde{Y}$ by

$$\langle y^*, \tilde{y}\rangle := \langle v^*, v\rangle + \langle w^*, w\rangle \text{ for } \tilde{y} = \tilde{S}v + \tilde{E}w. \tag{21}$$

To show continuity of y^* and thus $y^* \in \tilde{Y}^*$, let $\varepsilon > 0$ and choose $v \in U$, and $w \in Y_\infty$ such that $\tilde{y} = \tilde{S}v + \tilde{E}w$ and $\|v\|_U + \|w\|_{Y_\infty} \le \|\tilde{y}\|_{\tilde{Y}} + \varepsilon$. Then

$$|\langle y^*, \tilde{y}\rangle| = |\langle v^*, v\rangle + \langle w^*, w\rangle| \le \|v^*\|_{U^*}\|v\|_U + \|w^*\|_{Y_\infty^*}\|w\|_{Y_\infty} \le C(\|\tilde{y}\|_{\tilde{Y}} + \varepsilon).$$

Finally, setting $z^* := (0, y^*)$ we compute by (21) for all $x \in X$:

$$\langle \tilde{T}^*z^*, x\rangle = \langle z^*, \tilde{T}x\rangle = \langle 0, v\rangle + \langle y^*, \tilde{S}v + \tilde{E}w\rangle = \langle v^*, v\rangle + \langle w^*, w\rangle = \langle x^*, x\rangle,$$

and thus $\tilde{T}^*z^* = x^*$. □

It is clear from (19) that $\tilde{T}^*$ is injective, iff $\tilde{E}^*$ is injective. Then possible solutions of the equation $\tilde{T}^*z^* = x^*$ are unique.

4.2 Analysis of the transformed problem

Now we come to the core of our proof: the application of the sum-rule and the chain rule of subdifferential calculus (Theorem B.2) to our problem.

Proposition 4.4. *Suppose that the Assumptions 2.3 and 2.4 hold. For all $x \in X$ the following equation holds in X^*:*

$$\partial(F \circ T)(x) = T^*\partial j(Tx) + T^*\partial\iota_{\mathcal{U}}(Tx) + \partial(\iota_{\mathcal{Y}} \circ T)(x) + \partial(\iota_{\mathcal{K}} \circ T)(x). \tag{22}$$

Proof. Recall that $F = j + \iota_{\mathcal{Z}} = j + \iota_{\mathcal{U}} + \iota_{\mathcal{Y}} + \iota_{\mathcal{K}}$, and thus

$$\partial(F \circ T)(x) = \partial(j \circ T + \iota_{\mathcal{U}} \circ T + \iota_{\mathcal{Y}} \circ T + \iota_{\mathcal{K}} \circ T)(x). \tag{23}$$

We will apply Theorem B.2 to (23) three times to obtain (22) step by step. The interesting step, which only works in the space X, is the third one.

For the later application of Theorem B.2 we remark that the spaces X and Z are complete by Assumption 2.3*(i)* and Assumption 2.4*(i)* and all functions involved in our computations are convex and lower semi-continuous: either directly by assumption, or as a composition of a continuous mapping and a lower semi-continuous function. Thus it remains to show the crucial regularity condition (73) for each particular step.

Further will use the following simple identity: if $M \subset Z$, and $L : X \to Z$, then

$$L\,\mathrm{dom}(\iota_M \circ L) = L\{x \in X : Lx \in M\} = \{Lx \in Z : Lx \in M\} = \mathrm{ran}\,L \cap M. \tag{24}$$

We will need Assumption 2.4*(ii)*: existence of a Slater point $(\breve{u}, \breve{y})$. We may assume w.l.o.g. that $(\breve{u}, \breve{y}) = 0$. Otherwise, a simple shift by $(-\breve{u}, -\breve{y})$ can be performed to achieve this. This simplification is possible, because $(\breve{u}, \breve{y}) \in \mathcal{Z} \cap Z_\infty$ and it implies that $0 \in \mathcal{Z}$.

Step 1: splitting off the objective functional

First, we will show the regularity condition (73) for $g = j$, $f = \iota_{\mathcal{Z}} \circ T$, and $L = T$. By Framework 2.1*(iv)* $j : Z \to \mathbb{R}$ and thus $\operatorname{dom} j = Z$. By Assumption 2.3*(ii)* $\mathcal{Z}$ is non-empty, and by Proposition 4.1 $\mathcal{Z} \subset \mathcal{K} \subset \operatorname{ran} T$. Hence, by (24)

$$0 \in \operatorname{core}(\operatorname{dom} j - T \operatorname{dom}(\iota_{\mathcal{Z}} \circ T)) = \operatorname{core}(Z - \operatorname{ran} T \cap \mathcal{Z}) = Z,$$

and Theorem B.2 yields

$$\partial(F \circ T)(x) = T^* \partial j(Tx) + \partial\left(\iota_{\mathcal{Z}} \circ T\right)(x) \text{ in } X^*. \tag{25}$$

Step 2: splitting off the control constraints

Next, we will show (73) for $g = \iota_{\mathcal{U}}$, $f = \iota_{\mathcal{Y}} \circ T + \iota_{\mathcal{K}} \circ T$, and $L = T$ which reads

$$0 \in \operatorname{core}(\operatorname{dom} \iota_{\mathcal{U}} - T \operatorname{dom}(\iota_{\mathcal{Y}} \circ T + \iota_{\mathcal{K}} \circ T)). \tag{26}$$

First of all $\operatorname{dom} \iota_{\mathcal{U}} = \mathcal{U} \times Y$. Further, $T \operatorname{dom}(\iota_{\mathcal{Y}} \circ T + \iota_{\mathcal{K}} \circ T) = \operatorname{ran} T \cap (U \times \mathcal{Y}) \cap \mathcal{K}$ by (24) and $\mathcal{K} \subset \operatorname{ran} T$ by Proposition 4.1. So (26) reduces to

$$0 \in \operatorname{core}\left(\mathcal{U} \times Y - (U \times \mathcal{Y}) \cap \mathcal{K}\right). \tag{27}$$

To verify (27), let $z = (u, y) \in Z$. We have to find z_f and z_g such that $z_f \in (U \times \mathcal{Y}) \cap \mathcal{K}$, $z_g \in \mathcal{U} \times Y$ and for some $\lambda > 0$, $\lambda(z_f - z_g) = z$.

By Assumption 2.4*(iii)* there are $u_\infty \in U_\infty$, $u_{ad} \in \mathcal{U}$, and $\lambda_0 > 0$, such that $u = \lambda_0(u_{ad} + u_\infty)$. As stated in Framework 2.2*(iii)* $Su_\infty \in Y_\infty$. By Assumption 2.4*(ii)* (w.l.o.g. $(\breve{u}, \breve{y}) = 0$) there is $\tau > 0$ such that $y \in \mathcal{Y}$ for all $y \in Y_\infty$ with $\|y\|_{Y_\infty} \leq \tau$.

Let $\sigma := \min\{1, \tau / \|Su_\infty\|_{Y_\infty}\}$ and $z_f = (u_f, y_f) := -\sigma(u_\infty, Su_\infty)$. It follows $\|y_f\|_{Y_\infty} \leq \tau$ and $y_f = Su_f$, and thus $z_f \in (U \times \mathcal{Y}) \cap \mathcal{K}$. Set $\lambda := \lambda_0 / \sigma$ and $z_g = (u_g, y_g) := (\sigma u_{ad}, \lambda^{-1} y + y_f)$. Then $z_g \in \mathcal{U} \times Y$, because $\sigma \leq 1$ and $\mathcal{U}$ is convex with $0 \in \mathcal{U}$ w.l.o.g.. We obtain

$$\lambda(z_g - z_f) = (\lambda_0 \sigma^{-1}(u_g - u_f), \lambda(y_g - y_f)) = (\lambda_0(u_{ad} + u_\infty), (y + \lambda y_f) - \lambda y_f) = (u, y).$$

This shows (27) and we conclude by Theorem B.2:

$$\partial\left(\iota_{\mathcal{Z}} \circ T\right)(x) = T^* \partial \iota_{\mathcal{U}}(Tx) + \partial\left(\iota_{\mathcal{K}} \circ T + \iota_{\mathcal{Y}} \circ T\right)(x) \text{ in } X^*. \tag{28}$$

Step 3: separation of state constraints and equality constraints

Finally, we will show (73) for $g = \iota_{\mathcal{Y}} \circ T$, $f = \iota_{\mathcal{K}} \circ T$, and $L = Id_X$, which reads

$$0 \in \operatorname{core}\left(\operatorname{dom}(\iota_{\mathcal{Y}} \circ T) - \operatorname{dom}(\iota_{\mathcal{K}} \circ T)\right). \tag{29}$$

To verify (29) let $x = (v, w) \in X$. We have to find x_f and x_g such that $Tx_f \in \mathcal{K}$, $Tx_g \in \mathcal{Y}$ and for some $\lambda > 0$, $x = \lambda(x_f - x_g)$. By our Assumption 2.4*(ii)* (with $(\breve{u}, \breve{y}) = 0$ w.l.o.g.) there is $\tau > 0$, such that $y \in \mathcal{Y}$, for all $y \in Y_\infty$ with $\|y\|_{Y_\infty} \leq \tau$.

Let $\sigma := \tau / \left\|w\right\|_{Y_\infty}$ and set $x_f := (\sigma v, 0)$, $x_g := (0, -\sigma w)$, and $\lambda := \sigma^{-1}$. Then $Tx_f = \sigma(v, Sv) \in \mathcal{K}$, and $Tx_g = (0, -\sigma w) \in \mathcal{Y}$, because $\left\|\sigma w\right\|_{Y_\infty} \leq \tau$. Further, $\lambda(x_f - x_g) = (v, w) = x$. Hence, (29) holds and Theorem B.2 yields

$$\partial\left(\iota_{\mathcal{K}} \circ T + \iota_{\mathcal{Y}} \circ T\right)(x) = \partial(\iota_{\mathcal{K}} \circ T)(x) + \partial(\iota_{\mathcal{Y}} \circ T)(x) \text{ in } X^*. \tag{30}$$

Now (25), (28), and (30) yield (22). □

Next we study $\partial(\iota_{\mathcal{Y}} \circ T)(x)$ and $\partial(\iota_{\mathcal{K}} \circ T)(x)$ that appear in (22).

Lemma 4.5. *Let $f : \tilde{Z} \to \overline{\mathbb{R}}$ be independent of u, i.e., $f(z) = f(y)$. Then*

$$\partial(f \circ \tilde{T})(x) = \tilde{T}^* \partial f(\tilde{T}x) \quad \forall x \in X. \tag{31}$$

Proof. Consider $x \in X$, $(u, y) = z = \tilde{T}x \in \tilde{Z}$, and assume $f(\tilde{T}x) = f(y) < \infty$. Otherwise (31) holds trivially.

If $z^* \in \partial f(\tilde{T}x)$, then by definition of the subdifferential $\langle z^*, \hat{z} - \tilde{T}x\rangle \leq f(\hat{z}) - f(\tilde{T}x)$ for all $\hat{z} \in \tilde{Z}$ and in particular for all $\tilde{T}\hat{x} \in \operatorname{ran}\tilde{T}$. Thus,

$$\langle \tilde{T}^* z^*, \hat{x} - x\rangle = \langle z^*, \tilde{T}(\hat{x} - x)\rangle \leq f(\tilde{T}\hat{x}) - f(\tilde{T}x) = (f \circ \tilde{T})(\hat{x}) - (f \circ \tilde{T})(x) \quad \forall \hat{x} \in X.$$

Hence, $\tilde{T}^* z^* \in \partial(f \circ \tilde{T})(x)$ and thus $\tilde{T}^* \partial f(\tilde{T}x) \subset \partial(f \circ \tilde{T})(x)$.

For the reverse inclusion let $x^* \in \partial(f \circ \tilde{T})(x)$. We have to show existence of $z^* \in \partial f(\tilde{T}x)$ with $\tilde{T}^* z^* = x^*$. For this we will need Proposition 4.3*(iii)*.

Let $\delta x = (\delta v, \delta w) \in X$, with $\tilde{S}\delta v + \tilde{E}\delta w = 0$. Then $\tilde{T}(x + \delta x) = (u + \delta v, y + 0)$ and $f(\tilde{T}(x + \delta x)) = f(y) = f(\tilde{T}x)$ because f is independent of u. We conclude

$$\langle x^*, \delta x\rangle \leq (f \circ \tilde{T})(x + \delta x) - (f \circ \tilde{T})(x) = f(y) - f(y) = 0.$$

This holds also for $-\delta x$ and hence $\langle x^*, \delta x\rangle = 0$. Thus x^* vanishes for all δx that satisfy $\tilde{S}\delta v + \tilde{E}\delta w = 0$. By Proposition 4.3*(iii)* there is $y^* \in \tilde{Y}^*$, such that $z^* = (0, y^*)$ and $x^* = \tilde{T}^* z^*$.

To show that $z^* \in \partial f(\tilde{T}x)$, we have to verify $\langle z^*, \hat{z} - \tilde{T}x\rangle \leq f(\hat{z}) - f(\tilde{T}x)$ for all $\hat{z} = (\hat{u}, \hat{y}) \in \tilde{Z}$. Because $\hat{y} \in \tilde{Y} = \operatorname{ran}\tilde{S} + Y_\infty$, by definition of $\tilde{T}$ in (16) there exist $\bar{x}$ and $\bar{u}$, such that $\tilde{T}\bar{x} = \bar{z} := (\bar{u}, \hat{y})$. Since z^* and f do not depend on u, we compute $\langle z^*, \hat{z}\rangle = \langle y^*, \hat{y}\rangle = \langle z^*, \bar{z}\rangle$ and $f(\hat{z}) = f(\hat{y}) = f(\bar{z}) = f(\tilde{T}\bar{x})$. Thus, because $x^* \in \partial(f \circ \tilde{T})(x)$

$$\begin{aligned}\langle z^*, \hat{z} - \tilde{T}x\rangle &= \langle z^*, \bar{z} - \tilde{T}x\rangle = \langle z^*, \tilde{T}(\bar{x} - x)\rangle = \langle \tilde{T}^* z^*, \bar{x} - x\rangle \\ &= \langle x^*, \bar{x} - x\rangle \leq (f \circ \tilde{T})(\bar{x}) - (f \circ \tilde{T})(x) = f(\hat{z}) - f(\tilde{T}x).\end{aligned}$$

Hence, $z^* \in \partial f(\tilde{T}x)$ and thus $\tilde{T}^* \partial f(\tilde{T}x) \supset \partial(f \circ \tilde{T})(x)$. □

Clearly, $\iota_{\mathcal{Y}} \circ E_Z : \tilde{Z} \to \overline{\mathbb{R}}$ is independent of u, and thus Lemma 4.5 yields

$$\partial(\iota_{\mathcal{Y}} \circ T)(x) = \partial((\iota_{\mathcal{Y}} \circ E_Z) \circ \tilde{T})(x) = \tilde{T}^* \partial(\iota_{\mathcal{Y}} \circ E_Z)(\tilde{T}x) \quad \forall x \in X. \tag{32}$$

Lemma 4.6. *Suppose that Assumption 2.5 holds. Let* $\operatorname{dom} K := U \times \operatorname{dom} A_\infty \subset X$. *Then*

$$\begin{aligned} K : X \supset \operatorname{dom} K &\to R_\infty \\ (v, w) &\mapsto A_\infty w. \end{aligned}$$

is a densely defined, closed linear operator with closed range and $U = \ker K$. *Its adjoint operator* $K^* : R_\infty^* \supset \operatorname{dom} K^* \to X^*$ *is given by* $\operatorname{dom} K^* = \operatorname{dom} A_\infty^*$ *and*

$$\langle K^* r^*, x \rangle = \langle A_\infty^* r^*, w \rangle \quad \forall r^* \in \operatorname{dom} K^* \quad \forall x = (v, w) \in X. \tag{33}$$

It holds

$$\partial(\iota_{\mathcal{K}} \circ T)(x) = \operatorname{ran} K^*. \tag{34}$$

Proof. We will deduce the properties of K form the ones of A and A_∞ stated in Framework 2.1*(i)*, 2.2*(ii)*, and Assumption 2.5. We compute

$$\operatorname{graph}(K) = \{(v, w, r) \in U \times \operatorname{dom} A_\infty \times R_\infty : A_\infty w = r\} = U \times \operatorname{graph}(A_\infty).$$

By closedness of A_∞, $U \times \operatorname{graph}(A_\infty)$ is closed and thus K is closed. Density of $\operatorname{dom} K$ and closedness of $\operatorname{ran} K$ follows immediately from the corresponding assumptions on A_∞. Finally, A_∞ inherits injectivity from A and we conclude $U = \ker K$.

Let $r^* \in R_\infty^*$. Then $\langle r^*, Kx \rangle = \langle r^*, A_\infty w \rangle$ for all $x \in \operatorname{dom} K$. In particular $\langle r^*, K\cdot \rangle$ is continuous on $\operatorname{dom} K$, iff $\langle r^*, A_\infty \cdot \rangle$ is continuous on $\operatorname{dom} A_\infty$. By definition of adjoints (cf. Appendix A.2) this implies $\operatorname{dom} A_\infty^* = \operatorname{dom} K^*$ and (33).

Since $U = \ker K$ and T maps U onto $\mathcal{K}$ bijectively, we have $\iota_{\mathcal{K}} \circ T = \iota_U = \iota_{\ker K}$, and thus by (70) we conclude (34). □

4.3 Back-transformation

Finally, we transform (22) from X^* back to $\tilde{Z}^*$.

Proposition 4.7. *Suppose that Assumption 2.5 holds and let* $a^* \in \tilde{Y}^*$ *and* $r^* \in \operatorname{dom} A_\infty^* \subset R_\infty^*$.

(i) If $\tilde{E}^* a^* = A_\infty^* r^*$, *then for all* $u \in U_\infty$ *we conclude* $\langle \tilde{S}^* a^*, u \rangle = \langle r^*, Bu \rangle$ *and* $|\langle r^*, Bu \rangle| \leq C \|u\|_U$.

(ii) $z^* \in \tilde{Z}$ *is a solution of the equation* $\tilde{T}^* z^* = K^* r^*$, *if and only if there is* $a^* \in \tilde{Y}$, *such that* $\tilde{E}^* a^* = A_\infty^* r^*$ *and* $z^* = (-\tilde{S}^* a^*, a^*)$.

Proof. Let $a^* \in \tilde{Y}^*$, $\tilde{E}^* a^* = A_\infty^* r^* \in Y_\infty^*$ and $u \in U_\infty$. Then $\tilde{S} u = Su \in \operatorname{dom} A_\infty \subset Y_\infty$ by Framework 2.2*(iii)*, and we can write $\tilde{E}\tilde{S}u = \tilde{S}u \in Y_\infty$. Further, we use that A and A_∞ coincide on $\operatorname{dom} A$ and thus $Bu = ASu = A_\infty \tilde{S} u \in R_\infty$. We compute

$$\langle \tilde{S}^* a^*, u \rangle = \langle \tilde{E}^* a^*, \tilde{S} u \rangle = \langle A_\infty^* r^*, \tilde{S} u \rangle = \langle r^*, A_\infty \tilde{S} u \rangle = \langle r^*, Bu \rangle,$$

and thus by continuity of the operator $\tilde{S} : U \to \tilde{Y}$ (cf. Proposition 3.3):

$$|\langle r^*, Bu\rangle| = |\langle a^*, \tilde{S}u\rangle| \leq \|a^*\|_{\tilde{Y}^*} \|\tilde{S}u\|_{\tilde{Y}} \leq C\,\|u\|_U \quad \forall\, u \in U_\infty.$$

To show *(ii)* we observe that Lemma 4.6 yields $K^*r^* = (0, A_\infty^* r^*)$. Hence, the characterization of solutions $z^* = (-\tilde{S}^*a^*, a^*)$ follows directly from Proposition 4.3*(ii)*.
□

Theorem 4.8. *Suppose that the Assumptions 2.3-2.5 hold.* $z_{opt} = (u_{opt}, y_{opt}) \in Z$ *is a minimizer of* (P) *if and only if* $z_{opt} \in \mathcal{Z}$ *and the following nonlinear system of equations has a solution* $(j^*, y^*, u^*, a^*) \in Z^* \times \tilde{Y}^* \times U^* \times \tilde{Y}^*$*:*

$$E^*j_y^* + y^* + a^* = 0 \ \text{in}\ \tilde{Y}^* \tag{35}$$

$$j_u^* + u^* - \tilde{S}^*a^* = 0 \ \text{in}\ U^* \tag{36}$$

$$\langle y^*, \tilde{y} - y_{opt}\rangle \leq 0\ \forall \tilde{y} \in \tilde{\mathcal{Y}} \tag{37}$$

$$\langle u^*, u - u_{opt}\rangle \leq 0\ \forall u \in \mathcal{U} \tag{38}$$

$$j^* = (j_u^*, j_y^*) \in \partial j(z_{opt}) \tag{39}$$

$$\exists r^* \in \operatorname{dom} A_\infty^* \subset R_\infty^* : \tilde{E}^*a^* = A_\infty^* r^*. \tag{40}$$

In this case

$$\langle j_u^*, \delta u\rangle + \langle u^*, \delta u\rangle - \langle r^*, B\delta u\rangle = 0\ \forall \delta u \in U_\infty \subset U. \tag{41}$$

Proof. By Proposition 4.2 and Proposition 4.4, $z_{opt} = Tx_{opt}$ is a minimizer of (P) if and only if in X^*:

$$0 \in \partial(F \circ T)(x_{opt}) = T^*\partial j(z_{opt}) + T^*\partial\iota_{\mathcal{U}}(z_{opt}) + \partial(\iota_{\mathcal{Y}} \circ T)(x_{opt}) + \partial(\iota_{\mathcal{K}} \circ T)(x_{opt}).$$

Inserting (34), (32), and $T^* = (E_Z\tilde{T})^* = \tilde{T}^*E_Z^*$ (cf. (17) and (6)) we obtain

$$0 \in \tilde{T}^*E_Z^*\partial j(z_{opt}) + \tilde{T}^*E_Z^*\partial\iota_{\mathcal{U}}(z_{opt}) + \tilde{T}^*\partial(\iota_{\mathcal{Y}} \circ E_Z)(\tilde{T}x_{opt}) + \operatorname{ran} K^*.$$

This is equivalent to existence of $j^* \in \partial j(z_{opt})$, $u^* \in \partial\iota_{\mathcal{U}}(z_{opt})$, $y^* \in \partial(\iota_{\mathcal{Y}} \circ E_Z)(\tilde{T}x_{opt})$, and $r^* \in \operatorname{dom} K^*$, such that with $z^* := E_Z^*j^* + E_Z^*u^* + y^*$ the equation

$$0 = \tilde{T}^*z^* + K^*r^*.$$

is solvable for z^*. By Proposition 4.7 this is equivalent to existence of $a^* \in \tilde{Y}$ with $\tilde{E}^*a^* = -A_\infty^*r^*$, such that $z^* = -(-\tilde{S}^*a^*, a^*)$ and thus to solvability of the equation

$$0 = z^* + (-\tilde{S}^*a^*, a^*) = E_Z^*j^* + E_Z^*u^* + y^* + (-\tilde{S}^*a^*, a^*) \text{ in } \tilde{Z}^*. \tag{42}$$

Now $\iota_{\mathcal{Y}}$ is independent of u, and thus $y^* \in \tilde{Z}^*$, too, and thus actually $y^* \in \tilde{Y}^*$. Similarly, $u^* \in \tilde{Z}^*$ is independent of y and thus $E_Z^*u^* = u^* \in U^*$. Splitting (42) into its components in $\tilde{Y}^*$ and U^* yields (35) and (36) by $E_Z^*j^* = (j_u^*, E^*j_y^*)$ (cf. (6)). By (69), u^* and y^* satisfy (38), and (37), respectively. Hence, z_{opt} is a minimizer of (P), if and only if (35)-(40) is solvable, namely by j^*, u^*, y^*, and a^*. Finally, (41) follows from Proposition 4.7(i) and (36). □

Theorem 3.4 is now a special case of Theorem 4.8 if U_∞ is dense in U, as shown in Section 3. By reformulation we obtain from Theorem 4.8 another interesting result.

Corollary 4.9. *Suppose that Assumptions 2.3-2.5 hold.* $z_{opt} = (u_{opt}, y_{opt}) \in Z$ *is a minimizer of* (P) *if and only if* $z_{opt} \in \mathcal{Z}$ *and the following nonlinear system of equations has a solution* $(j^*, y^*, u^*, r^*) \in Z^* \times \tilde{Y}^* \times U^* \times R_\infty^*$*:*

$$E^* j_y^* + y^* + A_\infty^* r^* = 0 \text{ in } Y_\infty^* \tag{43}$$

$$S^* j_y^* + \tilde{S}^* y^* + j_u^* + u^* = 0 \text{ in } U^* \tag{44}$$

$$\langle y^*, \tilde{y} - y_{opt} \rangle \leq 0 \; \forall \tilde{y} \in \tilde{\mathcal{Y}} \tag{45}$$

$$\langle u^*, u - u_{opt} \rangle \leq 0 \; \forall u \in \mathcal{U} \tag{46}$$

$$j^* = (j_u^*, j_y^*) \in \partial j(z_{opt}), \quad r^* \in \operatorname{dom} A_\infty^*. \tag{47}$$

Proof. Solvability of (43)-(45) follows directly from solvability of (35)-(40) if we solve (35) for $a^* = -E^* j_y^* - y^*$ and insert this into (36).

For the converse, set $a^* := -E^* j_y^* - y^* \in \tilde{Y}^*$. By (43), $\tilde{E}^* a^* = A_\infty^* r^*$, and by (44), $\tilde{S}^* a^* = j_u^* + u^* \in U^*$. It is now easy to verify that (35)-(40) hold. □

5 Applications

Finally we illustrate the application of our abstract results to optimal control problems and describe the relation of our theory to known results.

5.1 Relation to known results

Basically two types of known results are available. Both approaches consider the reduced problem (with differentiable j)

$$\min_{u \in U} j(u, Su) + (\iota_{\mathcal{Y}} \circ S)(u)$$

and start by application of Theorem B.2, (or a similar result, e.g., [21]) to show

$$0 \in j_u(u_{opt}) + S^* j_y(Su_{opt}) + \partial(\iota_{\mathcal{Y}} \circ S)(u_{opt}). \tag{48}$$

The first type uses the assumption that the control-to-state mapping $S : U \to Y$ is continuous for the choice $Y = Y_\infty$. Exemplary works are [5, 6, 1]. Under a Slater condition, Theorem B.2 can now be used to show $\partial(\iota_{\mathcal{Y}} \circ S)(u_{opt}) = S^* \partial \iota_{\mathcal{Y}}(Su_{opt})$. This yields existence of a Lagrange multiplier $y^* \in \partial \iota_{\mathcal{Y}}(Su_{opt}) \subset Y_\infty^*$. The most difficult task is now to interpret S^* via an adjoint PDE, which yields additional regularity assertions on the optimal solution. In the case of pointwise state constraints Y_∞ is a space of continuous functions, and thus its dual has a representation as a space of regular measures.

This type of results is, at least in an abstract sense, covered by Theorem 3.4 as a special case, namely the “most regular” extreme case: $U_\infty = U$. Then $\tilde{Y} \cong Y_\infty$, and $\tilde{Y}^* \cong Y_\infty^*$. So Theorem 3.4 and Theorem 4.8 may be interpreted as a generalizations of these results.

The second approach corresponds to the “trivial” extreme case of Theorem 4.8, namely $(Z_\infty, R_\infty) = (\{0\}, \{0\})$. Hence, there are no restricting topological assumptions and no Slater condition. But this also means that the structure of the problem is lost: pointwise *equality* constraints are not distinguished from *inequality* constraints.

The analysis starts with (48), uses bijectivity of $S : U \to \operatorname{ran} S$ and reformulates $\partial(\iota_{\mathcal{Y}} \circ S) = S^*(S^{-*}\partial(\iota_{\mathcal{Y}} \circ S))$. Then $y^* \in S^{-*}\partial(\iota_{\mathcal{Y}} \circ S)$ is called a Lagrange multiplier. In contrast to the first approach no additional structural information on the optimal solution can be shown, compared to conclusions drawn from coercivity of the functional. This type of dual variable can, for example, be observed as the limit object of the regularization path considered in [12].

5.2 Elliptic Differential Operators as Closed Operators

Let Ω be a smoothly bounded domain of $\mathbb{R}^d$, $\kappa \in C(\Omega, \mathbb{R}^{d\times d})$, $a \in L_\infty(\Omega; \mathbb{R})$. Assume that κ is symmetric and uniformly positive definite and $0 \neq a \geq 0$.

First we consider the following class of elliptic differential operators in the weak form:

$$\begin{aligned} A &: H^1(\Omega) \to (H^1(\Omega))^* \\ \langle Ay, v\rangle &= \int_\Omega \langle \kappa\nabla y, \nabla v\rangle + ayv\,dt \quad \forall v \in H^1(\Omega). \end{aligned} \tag{49}$$

The Lax Milgram lemma asserts that A has a continuous inverse $A^{-1} : (H^1(\Omega))^* \to H^1(\Omega)$, which is still continuous as a mapping $A^{-1} : (H^1(\Omega))^* \to L_2(\Omega)$, because $H^1(\Omega)$ is continuously embedded into $L_2(\Omega)$. Hence,

$$A : L_2(\Omega) \supset H^1(\Omega) \to (H^1(\Omega))^* \tag{50}$$

is continuously invertible.

To define an operator $A_\infty : C(\overline{\Omega}) \supset \operatorname{dom} A_\infty \to R$ we have to employ advanced regularity results, which can be found in the literature in many variants. A concise account on regularity theory is [2, Section 9]. For our class of problems [2, Theorem 9.3] states that for $\infty > q > d$, and $q' = q/(q-1)$ the restricted mapping

$$A : W^{1,q}(\Omega) \leftrightarrow (W^{1,q'}(\Omega))^*$$

is an isomorphism. By the Sobolev embedding theorem $W^{1,q}(\Omega) \hookrightarrow C(\overline{\Omega})$ densely. Setting $\operatorname{dom} A_\infty := W^{1,q}(\Omega)$, we conclude by Lemma A.1 closedness and bijectivity of

$$A_\infty : C(\overline{\Omega}) \supset \operatorname{dom} A_\infty \to (W^{1,q'}(\Omega))^*. \tag{51}$$

As for the adjoint of A_∞, we have

$$\begin{aligned} A_\infty^* &: W^{1,q'}(\Omega) \supset \operatorname{dom} A_\infty^* \to M(\overline{\Omega}) \\ \langle y, A_\infty^* p\rangle &= \int_\Omega \langle \kappa \nabla y, \nabla p\rangle + ayp\, dt \quad \forall p \in W^{1,q}(\Omega). \end{aligned} \tag{52}$$

Theorem 5.1. *For each $\mu \in M(\overline{\Omega})$ the equation*

$$\int_\Omega \langle \kappa \nabla y, \nabla p\rangle + ayp\, dt = \int_\Omega p\, d\mu \quad \forall y \in W^{1,q}(\Omega) \tag{53}$$

has a unique solution p, and $\|p\|_{W^{1,q'}(\Omega)} \le C\, \|\mu\|_{M(\overline{\Omega})}$.

Proof. Since A_∞ is bijective, Theorem A.2 yields bijectivity of A_∞^*. Because all adjoint operators are closed, Lemma A.1 (essentially the open mapping theorem) then asserts continuous invertibility of A_∞^*. □

The regularity requirements on $\partial\Omega$, and the coefficients can be weakened considerably. The case of discontinuous κ is particularly delicate, and has been analysed in [1]. Yet, A_∞ can still be declared as a closed, bijective operator, and our abstract theory can be applied. All the information we need is stated in [1, Theorem 2], which asserts continuity of

$$A^{-1} : (W^{1,q'}(\Omega))^* \to C(\overline{\Omega}) \cap H^1(\Omega).$$

However, opposed to the regular case, $\operatorname{dom} A_\infty := \operatorname{ran} A^{-1}$ cannot be characterized as a Sobolev space. Despite of this lack of information, we still can conclude that

$$A_\infty^* : W^{1,q'}(\Omega) \supset \operatorname{dom} A_\infty^* \to M(\overline{\Omega})$$

is bijective.

The difficulties encountered and solved in [1] now appear merely as the problem of finding a convenient *representation* of the functional $A_\infty^* p$ via integrals. The difficulty is that the integral on the left hand side in (53) is not necessarily well defined for all $y \in \operatorname{dom} A_\infty$. This is a similar situation as in Section 3.4. Here [1] resort to an additional uniqueness criterion, which is connected to the formula of partial integration and to very weak solutions.

5.3 State constrained boundary control

As an example and illustration of our abstract results we consider the following optimal control problem on a smoothly bounded domain $\Omega \subset \mathbb{R}^3$ and with coefficients κ and a as defined in the above section.

$$\min_{(u,y)\in U\times Y} j(y,u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Gamma)} \tag{54}$$

subject to the state equation in the weak formulation

$$\int_\Omega \langle \kappa \nabla y, \nabla v\rangle + ayv\, dt = \int_\Gamma u \cdot \gamma(v)\, dt \quad \forall v \in H^1(\Omega), \tag{55}$$

(γ is the boundary trace operator), a pointwise state constraint, and a control constraint

$$\underline{y} \le y \quad \text{a.e. in } \Omega, \quad \underline{u} \le u \quad \text{a.e. in } \Gamma. \tag{56}$$

We assume that $\underline{y}$ is continuous and that there is a Slater point $(\breve{u}, \breve{y}) \in L_{s_0}(\Gamma) \times C(\overline{\Omega})$ with $s_0 > 2$ that satisfies the state equation, the control constraint, and $\breve{y} - \underline{y} \ge \tau > 0$ a.e..

A class of control problems similar to our example has been analysed in [6, 1], however under the assumption that $u_{opt} \in L_s(\Gamma)$ for some $s > 2$ and thus $y_{opt} \in C(\overline{\Omega})$. This can only be guaranteed a-priori by bilateral L_∞-control constraints, or by a functional with stronger coercivity properties. In our setting, the optimal state may be unbounded and discontinuous.

Existence of an optimal solution $(u_{opt}, y_{opt}) \in L_2(\Gamma) \times L_2(\Omega)$ is standard. Uniqueness follows from the strict convexity of j. By convexity of the problem, (u_{opt}, y_{opt}) also minimizes our problem with the modification that $\underline{u} \le u$ is only required on the active set

$$\mathcal{A} := \{t \in \Gamma : u_{opt} = \underline{u}\},$$

which is defined up to a set of measure zero.

Before stating our optimality conditions, let us first describe, how this problem fits into our abstract setting. Choose $\infty > q > 3$ and $s_0 > s > 2$ such that for $s' = s/(s-1)$ and $q' = q/(q-1)$ the boundary trace operator γ maps $W^{1,q'}(\Omega)$ into $L_{s'}(\Gamma)$. This does not hold for $q > 3$ and $s' = s = 2$.

For Framework 2.1 define $U := L_2(\Gamma \setminus \mathcal{A}) \times L_s(\mathcal{A})$, $Y := L_2(\Omega)$ and $R := (H^1(\Omega))^*$. Since $\underline{u} \in L_\infty(\Gamma)$, clearly $u_{opt} \in U$ and $U^* \cong L_2(\Gamma \setminus \mathcal{A}) \times L_{s'}(\mathcal{A})$. Then, A as defined in (50) is continuously invertible. Since the boundary trace operator $\gamma : H^1(\Omega) \to U^*$ is continuous (it is even continuous onto $L_2(\Gamma) \hookrightarrow U^*$), the following representation of its adjoint operator

$$\begin{aligned} B &: U \to R \\ u &\mapsto \langle Bu, v\rangle = \int_\Gamma u\gamma(v)\, ds \quad \forall v \in H^1(\Omega) \end{aligned} \tag{57}$$

is continuous, too, and thus $S := A^{-1}B$ is continuous.

For Framework 2.2 define $U_\infty := L_s(\Gamma)$, $Y_\infty := C(\overline{\Omega})$, $R_\infty := (W^{1,q'}(\Omega))^*$. Choose A_∞ as in (51). Clearly, A_∞ and A coincide for $y \in \operatorname{dom} A_\infty = W^{1,q}(\Omega)$. By our choice of s, γ maps $W^{1,q'}(\Omega)$ into $L_{s'}(\Gamma)$ and thus by duality, B maps $U_\infty = L_s(\Gamma)$ into $R_\infty = (W^{1,q'}(\Omega))^*$. Because A_∞ is bijective, $SU_\infty = A_\infty^{-1} BU_\infty \in \operatorname{dom} A_\infty$.

We can now define the tailored space $\tilde{Y} = \operatorname{ran} S + Y_\infty = \operatorname{ran} S + C(\overline{\Omega})$ as in Definition 3.2. It contains all solutions of the Neumann boundary value problem and all continuous functions.

Theorem 5.2. *Let $(u_{opt}, y_{opt}) \in U \times Y$ be the optimal solution of the problem* (54)-(56)*. Then the following assertion holds.*

There exist $p \in W^{1,q'}(\Omega)$, $y^* \in \tilde{Y}$ with representation $0 \leq \mu \in M(\overline{\Omega})$, and $0 \leq u^* \in L_{s'}(\mathcal{A})$ that satisfy

$$\int_\Omega \varphi(y_{opt} - y_d)\, dt + \int_\Omega \langle \kappa \nabla \varphi, \nabla p\rangle + a\varphi p\, dt - \int_{\overline{\Omega}} \varphi\, d\mu = 0 \quad \forall\, \varphi \in W^{1,q}(\Omega) \tag{58}$$

$$\alpha u_{opt} - u^* - \gamma(p) = 0 \quad \text{a.e. in } \Gamma \tag{59}$$

$$\langle y^*, \delta y\rangle \geq 0 \quad \forall 0 \leq \delta y \in \tilde{Y} \tag{60}$$

$$\langle y^*, y_{opt} - \underline{y}\rangle = 0. \tag{61}$$

Equations (60) *and* (61) *have a representation of the form* (12)*.*

If in converse the system (58)-(61) *is solvable for given feasible (u_{opt}, y_{opt}), then (u_{opt}, y_{opt}) is the optimal solution of the problem* (54)-(56)*.*

Proof. To apply Theorem 3.4 and derive first order optimality conditions we have to verify the Assumptions 2.3-2.5. For this purpose choose $q > 3$ and $s > 2$ as stated above.

Defining $\mathcal{Y}$ via (56) and $\mathcal{U}$ via $\underline{u} \leq u$ on $\mathcal{A}$, Assumptions 2.3 are easily verified, and j is even Gâteaux differentiable with $j'(u_{opt}, y_{opt}) = (\alpha u_{opt}, y_{opt} - y_d)$. Existence of a Slater point $\breve{y} - \underline{y} \geq \tau > 0$ has been assumed in the statement of our problem, and by our choice $Y_\infty = C(\overline{\Omega})$ this directly translates into Assumption 2.4*(ii)*. Assumption 2.4*(iii)* follows from our particular construction of U, which is restricted to L_s on the active set. Clearly, all spaces involved are complete, and as shown in Section 5.2, A_∞ as defined in (51) satisfies Assumption 2.5.

Clearly, U_∞ is dense in U. Hence, Theorem 3.4 can be applied, which yields solvability of (7)-(10). By Proposition 3.5 y^* has a representation as a measure μ.

Now (58) and (59) are equivalent to (7) (via the definition of A^*) and (8) (via the fact that (8) is an equation in U^*, which is an L_p-space here). The remaining positivity assertions on y^* and u^* are equivalent to (9) and (10) by (71) and (72). □

Remark 5.3. There are several remarks in order:

(i) Observe that (58) is really the weak form of a PDE with a measure right hand side. This holds, because all test functions $\varphi \in W^{1,q}(\Omega) \hookrightarrow C(\overline{\Omega})$ are continuous.

(ii) As usual, our result holds for all small $q > d$. So $p \in \bigcap_{q' < d/(d-1)} W^{1,q'}(\Omega)$. This is the same regularity assertion as in the known cases. Even more, (59) in combination with the trace theorem shows that p has got a little bit more regular boundary trace: $\gamma(p)|_{\Gamma \setminus \mathcal{A}} \in L_2 \cap W^{1-1/q',q'}$, because $L_2 \not\hookrightarrow W^{1-1/q',q'}$.

(iii) From our result we can derive additional regularity properties for u_{opt}. In particular, (59) yields $u_{opt}|_{\mathcal{A}} \in L_\infty$ and $u_{opt}|_{\Gamma \setminus \mathcal{A}} \in L_2 \cap W^{1-1/q',q'}$.

(iv) Notice, how Assumption 2.4*(iii)* necessitates that the active set $\mathcal{A}$ is taken into account for the construction of U. The too large space $U = L_2(\Gamma)$ would yield $\lambda \in L_2(\mathcal{A})$, which is "too good to be true". The too small space $U = L_s(\Gamma)$ would not contain all possible candidates for a minimizer.

(v) Bilateral state constraints can be treated similarly if they yield a bounded (but possibly discontinuous) optimal state. We obtain a Lagrange multiplier $y^* \in \tilde{Y}^*$ and a representation $\mu \in M(\overline{\Omega})$ for the combined constraints. We can split y^* into two positive functionals in $L_\infty(\Omega)^*$ by Theorem B.2 and μ into two positive measures by the Jordan decomposition. It is unclear, however, if the summands are in $\tilde{Y}^*$ again. This does not affect the adjoint equation, because the test functions are continuous.

5.4 Missing density

In the following example U_∞ is not dense in U. As a consequence Theorem 3.4 cannot be applied, and we will show that indeed Lagrange multipliers cannot be represented as measures. Consider the unit ball in $\mathbb{R}^3$ and the following control problem with a one dimensional control $u \in \mathbb{R}$:

$$\min_{y \in H_0^1(\Omega), u \in \mathbb{R}} \frac{1}{2}|u+1|^2 \quad \text{s.t.} \quad \int_\Omega \langle \nabla y, \nabla v \rangle \, dt = \int_\Omega \langle f \cdot u, v \rangle \, dt, \qquad y \geq -1;$$

For appropriate choice of σ set $f = |t|^{-\sigma}$, such that y is an unbounded rational function with a pole at the origin, but still in $H_0^1(\Omega)$.

Clearly, for every $u < 0$ we have $\lim_{t \to 0} y(t) = -\infty$, which implies violation of the boundary conditions. Hence, our problem attains the minimum at $u_{opt} = 0$ and thus $j_u(u_{opt}) = 1$. Without state constraints, the minimum would have been $u = -1$, and thus the state constraints are strongly active. But $\operatorname{ess\,inf}_{t \in \Omega} y - (-1) = 1 > 0$ and the active constraint set is empty. Consequently, any positive element of $M(\overline{\Omega})$ or $L_\infty(\Omega)^*$ that satisfies a complementarity condition (72) must be zero.

Still, Corollary 4.9 can be applied. Since Su is unbounded for every $u \neq 0$, we have to choose $U_\infty = \{0\}$, which is clearly not dense in $U = \mathbb{R}$. Further, we choose $Y_\infty = \{0\}$ and $R_\infty = \{0\}$. Then $\tilde{Y} = \operatorname{ran} S$. Corollary 4.9 yields the optimality system (43)-(44). Equation (43) vanishes and (44) reads in our case $S^* y^* + j_u^* = 0$ in $U^* = \mathbb{R}$. Hence, $y^* \in \tilde{Y}$ is of the following form: let $\tilde{Y} \ni y = Su$, then $\langle y^*, y \rangle = \langle -j_u, u \rangle$. This and (45) yield the condition $\langle j_u, u - u_{opt} \rangle \geq 0$ if $Su \geq -1$, or equivalently, if $u \geq 0$. This holds, if and only if $u_{opt} = 0$ and thus characterizes the optimal solution.

This illustrates nicely how y^* behaves in the absence of a Slater point. It is rather the dual variable for an implicit control constraint ($u \geq 0$), than for the state constraint.

6 Conclusion and Outlook

We have presented a new technique for the analysis of state constrained optimal control problems. It allows to exploit Slater conditions in cases, where this seemed impossible. Abstract first order optimality conditions were derived, and their convenient application to control problems was demonstrated. Under a density assumption the Lagrange multipliers for the state constraints have a representation via measures.

While we have answered one question, many other theoretical and practical questions arise. First of all, the application of our results to various classes of optimal control problems may be explored systematically. In particular the extension to nonlinear, non-convex problems should be explored, and the consequences of these results to second order optimality conditions. Equally important is the analysis of algorithms. It will be interesting to study the convergence behaviour of infeasible regularization methods (cf. e.g. [12, 14]) in the light of our new results, and it is very likely that barrier methods in function space (cf. e.g. [17]) can also be analysed in this setting. For that, the indicator function $\iota_{\mathcal{Y}}$ may be simply replaced by an appropriate barrier function $b_{\mathcal{Y}}$. Finally, the construction of discretization schemes and their analysis remains as a challenging topic, because L_∞-error estimates for the state will not be available in general.

References

[1] J.-J. Alibert and Raymond J.-P. Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls. *Numer. Funct. Anal. and Optimization*, 3&4:235–250, 1997.

[2] H. Amann. Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems. In H.J. Schmeisser and H. Triebel, editors, *Function Spaces, Differential Operators and Nonlinear Analysis.*, pages 9–126. Teubner, Stuttgart, Leipzig, 1993.

[3] R.B. Ash. *Measure, Integration and Functional Analysis.* Academic Press, 1972.

[4] J.M. Borwein and Q.J. Zhu. *Techniques of Variational Analysis.* CMS Books in Mathematics. Springer, 2005.

[5] E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM J. Control Optim.*, 24(6):1309–1318, 1986.

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

[7] E. Casas. Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations. *SIAM J. Control and Optimization*, 35:1297–1327, 1997.

[8] E. Casas and L.A. Fernández. Optimal Control of Semilinear Elliptic Equations with Pointwise Constraints on the Gradient of the State. *Appl. Math. Optim.*, 27:35–56, 1993.

[9] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems.* Number 28 in Classics in Applied Mathematics. SIAM, 1999.

[10] D.H. Fremlin. *Toplogical Riesz Spaces and Measure Theory.* Cambridge University Press, 1974.

[11] S. Goldberg. *Unbounded Linear Operators.* Dover Publications, Inc., 1966.

[12] M. Hintermüller and K. Kunisch. Feasible and non-interior path-following in constrained minimization with low multiplier regularity. *SIAM J. Control Optim.*, 45(4):1198–1221, 2006.

[13] V. Lykina, S. Pickenhain, and M. Wagner. Different interpretations of the improper integral objective in an infinite horizon control problem. *J. Math. Anal. Appl.*, 340:498 – 510, 2008.

[14] C. Meyer, F. Tröltzsch, and A. Rösch. Optimal control problems of PDEs with regularized pointwise state constraints. *Computational Optimization and Applications*, 33:206–228, 2006.

[15] S.M. Robinson. Regularity and stability for convex multivalued functions. *Math. Oper. Res.*, 1:130–143, 1976.

[16] R.T. Rockafellar. *Conjugate duality and optimisation.* SIAM Publications, 1974.

[17] A. Schiela. Barrier methods for optimal control problems with state constraints. ZIB Report 07-07, Zuse Institute Berlin, 2007.

[18] A. Schiela. Optimality conditions for convex state constrained optimal control problems with discontinuous states. ZIB Report 07-35, Zuse Institute Berlin, 2007.

[19] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen. Theorie, Verfahren und Anwendungen.* Vieweg, 2005.

[20] D. Werner. *Funktionalanalysis.* Springer, 3$^{\text{rd}}$ edition, 2000.

[21] J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optimization*, 5:49–62, 1979.

A Tools from the theory of unbounded operators

We give a brief introduction to some basic concepts of the theory of unbounded operators that we have used in our work. Unbounded operators possess a rich theory that generalizes the theory of continuous operators in many respects, while retaining a large number of important results. For a detailed exposition we refer to [11], but most textbooks of functional analysis contain an introduction to unbounded operators.

Consider normed spaces Y and R. An unbounded operator

$$A : Y \supset \operatorname{dom} A \to R$$

is usually not defined everywhere, but has got a *domain of definition* $\operatorname{dom} A$. We say that A is *injective*, *surjective*, or *bijective*, if the mapping $A : \operatorname{dom} A \to R$ has this algebraic property. If $\operatorname{dom} A$ is dense in Y, then A is called *densely defined.*

For example, if A is a differential operator and Y is some function space, then $\operatorname{dom} A$ may be a subspace of functions that are differentiable in a suitable sense. The distinction between Y (which yields the topological structure) and $\operatorname{dom} A$ (which yields the algebraic structure) allows for additional flexibility when it comes to choosing an analytical framework for the problem under consideration. In an optimal control setting this allows us to consider problems with differential operators, using a topology that is suited for state constrained problems, e.g. the topology of $C(\overline{\Omega})$.

We will use unbounded operators mainly to formulate our results in a way that is more directly applicable to PDE constrained optimal control problems than results stated in terms of S and S^*. In particular, the adjoint PDE follows instantly from our abstract optimality conditions. We will demonstrate this in Section 5.

A.1 Closed operators

A standard regularity assumption for unbounded operators is *closedness.* A is called *closed* if $\operatorname{dom} A \supset y_k \to y$ and $Ay_k \to r$ imply $y \in \operatorname{dom} A$ and $Ay = r$. If A is closed, Y and R are complete, and $\operatorname{dom} A = Y$, then A is continuous by the well known closed graph theorem. Closedness has also a geometrical interpretation: A is closed, if and only if

$$\operatorname{graph}(A) := \{(y, r) \in \operatorname{dom} A \times R : Ay = r\} \subset Y \times R \tag{62}$$

is closed in $Y \times R$. Continuous operators that are defined on the whole domain space are closed, because $y_k \to y$ already implies $Ay_k \to Ay$. Closed operators have closed kernels, which follows from considering sequences with $Ay_k = 0$.

Lemma A.1. *Let Y and R be normed spaces and let $A : Y \supset \operatorname{dom} A \to R$ be a linear operator. If A possesses a continuous inverse $A^{-1} : R \to Y$, then A is closed. If Y and R are complete, then each closed bijective linear operator has a continuous inverse.*

Proof. Clearly, bijectivity is equivalent to existence of an inverse A^{-1}. If A^{-1} is continuous and defined on all of R, its graph is closed and

$$\operatorname{graph}(A) = \{(y,r) : Ay = r\} = \{(y,r) : y = A^{-1}r\} = \operatorname{graph}(A^{-1}).$$

So A is closed. If, in converse, A is closed and Y and R are complete, then the application of the open mapping theorem (cf. e.g.[20, Satz IV.4.4]) for closed operators yields continuity of A^{-1}. □

A.2 Adjoints of densely defined operators

Let us recapitulate the definition of the adjoint of a densely defined operator $A : Y \supset \operatorname{dom} A \to R$, which generalizes the adjoint of a continuous operator. For a normed space R we denote by R^* its dual, equipped with the canonical norm, and by $\langle \cdot, \cdot \rangle$ the dual pairing. Define

$$\operatorname{dom} A^* := \{r^* \in R^* : \text{ the linear functional } \langle r^*, A\cdot \rangle \text{ is continuous on } \operatorname{dom} A\}\,.$$

If $r^* \in \operatorname{dom} A^*$, then $\langle r^*, A\cdot \rangle$ has a unique continuous extension to a functional $y^* = A^*r^* \in Y^*$, because it is continuous on the dense subset $\operatorname{dom} A \subset Y$. This yields the definition of $A^* : R^* \supset \operatorname{dom} A^* \to Y^*$, and the relation

$$\langle A^*r^*, y \rangle = \langle r^*, Ay \rangle \quad \forall\, y \in \operatorname{dom} A\, \forall\, r^* \in \operatorname{dom} A^*.$$

In particular, $\operatorname{dom} A^*$ is canonically defined and depends on the topology of Y and R. Adjoint operators are always closed by [11, Theorem II.2.6].

The following theorem establishes relations between a densely defined operator and its adjoint. In spite of its unconspicuous appearance, (66) is a deep and important existence result. For a normed space X, let $U \subset X$ and $V^* \subset X^*$. We define their “orthogonal” complements as following:

$$U^\perp := \{x^* \in X^* : \langle x^*, x \rangle = 0\ \forall x \in U\} \tag{63}$$

$$V^*_\perp := \{x \in X : \langle x^*, x \rangle = 0\ \forall x^* \in V^*\}. \tag{64}$$

Theorem A.2 (Closed Range Theorem)**.** *Let Y, R be normed spaces. Assume that $A : Y \supset \operatorname{dom} A \to R$ is densely defined. Then*

$$\overline{\operatorname{ran} A} = (\ker A^*)_\perp. \tag{65}$$

In particular, if A has dense range, then A^* is injective. If additionally Y and R are complete and A is closed with closed range, then

$$\operatorname{ran} A^* = (\ker A)^\perp. \tag{66}$$

In particular, if A is injective, then A^* is surjective.

Proof. Equation (65) follows from [11, Theorem II.3.7]. If A has dense range, then $\overline{\operatorname{ran} A} = R = (\ker A^*)_\perp$ which implies $\ker A^* = \{0\}$, and A^* is injective. Equation (66) follows from [11, Theorem IV.1.2]. In particular, if A is injective with closed range, then $\operatorname{ran} A^* = (\ker A)^\perp = \{0\}^\perp$, hence A^* is surjective (here, the hypothesis of closed range cannot be dispensed with). □

B Tools from convex analysis

We will introduce some basic concepts and tools from convex analysis. For more details on convex analysis we refer to [9, Chapter I] or [4, Chapter 4].

It is customary in convex analysis to consider *extended real valued functions*

$$f : X \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}.$$

This makes it possible to consider constrained and unconstrained optimization problems in one framework, by setting $f = \infty$ for infeasible points. The set of points $\operatorname{dom} f$, where f takes a finite value is called the *domain* of f. Apart from *convexity*, standard assumptions on f are *lower semi-continuity* (i.e. the sets $\{x \in X : f(x) \leq \alpha\}$ are closed for all $\alpha \in \overline{\mathbb{R}}$) and *properness*: $\operatorname{dom} f \neq \emptyset$.

The *indicator function* ι_M of a set $M \subset X$ is defined by

$$\iota_M(x) = \left\{ \begin{array}{ccl} 0 & : & x \in M \\ \infty & : & \text{otherwise.} \end{array} \right. \tag{67}$$

It is convex and lower semi-continuous if and only if M is convex and closed, respectively, and $\operatorname{dom} \iota_M = M$. It follows from the definition that $\iota_{M_1} + \iota_{M_2} = \iota_{M_1 \cap M_2}$.

In convex analysis the usual differentiability concept is replaced by subdifferentiability. The *subdifferential* $\partial f(x)$ of $f : X \to \overline{\mathbb{R}}$ at a point $x \in \operatorname{dom} f$ is the set of all $x^* \in X^*$, for which the relation $\langle x^*, \hat{x} - x\rangle \leq f(\hat{x}) - f(x)$ holds for all $\hat{x} \in X$. If $f(x) = \infty$, then $\partial f(x)$ is defined to be the empty set.

If f is convex and Gâteaux differentiable at x with derivative $f'(x)$, then $\partial f(x) = \{f'(x)\}$ (cf. [9, Proposition I.5.3]). The following simple relation is a generalization of Fermat's principle:

$$0 \in \partial f(x_{opt}) \Leftrightarrow \langle 0, x - x_{opt}\rangle \leq f(x) - f(x_{opt}) \; \forall x \in X \Leftrightarrow f(x_{opt}) \leq f(x) \; \forall x \in X. \tag{68}$$

Hence, minimizers x_{opt} of f are characterized by $0 \in \partial f(x_{opt})$.

Lemma B.1. *Let X be a normed space and $x \in M \subset X$. The subdifferential $\partial\iota_M(x)$ is the set of all $x^* \in X^*$, which satisfy*

$$\langle x^*, \hat{x} - x\rangle \leq 0 \quad \forall \hat{x} \in M. \tag{69}$$

If M is a linear subspace of X, then $\partial\iota_M(x) = M^\perp$. If X and R are Banach spaces and $A : X \supset \operatorname{dom} A \to R$ is a closed, densely defined linear operator with closed range, then

$$\partial\iota_{\ker A}(x) = \operatorname{ran} A^* \; \forall\, x \in \ker A. \tag{70}$$

If X is a partially ordered normed space, $\underline{x} \in X$, and $M = \{x \in X : x \geq \underline{x}\}$, then $\partial\iota_M(x)$ is the set of all $x^* \in X$, which satisfy

$$\langle x^*, \delta x\rangle \leq 0 \quad \forall \delta x \geq 0, \tag{71}$$

$$\langle x^*, x - \underline{x}\rangle = 0. \tag{72}$$

Proof. If $x^* \in \partial\iota_M(x)$, then $\langle x^*, \hat{x} - x\rangle \leq \iota_M(\hat{x}) - \iota_M(x) = 0\,\forall \hat{x} \in X$, and in particular for all $\hat{x} \in M$. This is (69). If x^* satisfies (69), then $\iota_M(\hat{x}) = \infty$ for $\hat{x} \notin M$ yields $\langle x^*, \hat{x} - x\rangle \leq \iota_M(\hat{x}) - \iota_M(x)\,\forall \hat{x} \in X$, and thus $x^* \in \partial\iota_M(x)$.

If M is a linear subspace of X and $x \in M$, then (69) holds for $\delta x = \pm(\hat{x}-x) \in M$, which yields $\partial\iota_M(x) = M^\perp$. This, setting $M = \ker A$, together with (66) yields (70).

If $x \in M$, and $\delta x \geq 0$, then also $x + \delta x \in M$, and (69) yields (71). The choice $\delta x = \underline{x} - x$, yields $x + \delta x = \underline{x} \in M$ and $x - \delta x = 2x - \underline{x} \geq x \in \mathcal{Y}$, hence $\langle x^*, \pm\delta x\rangle \leq 0$, which yields (72). For the converse let $x, \hat{x} \in M$, and let $x^* \in X^*$ satisfy (71) and (72). Define $\delta x_M := x - \underline{x}$. Then (72) yields $\langle x^*, \delta x_M\rangle = 0$. Moreover, $\delta\hat{x} := (\hat{x} - x) + \delta x_M = \hat{x} - \underline{x} \geq 0$, because $\hat{x} \in M$. Thus $\langle x^*, \hat{x} - x\rangle = \langle x^*, \delta\hat{x} - \delta x_M\rangle = \langle x^*, \delta\hat{x}\rangle \leq 0$ by (71). Hence (69) holds, and $x \in \partial\iota_M(x)$. □

Apart from the closed range theorem (Theorem A.2) we will use a second deep existence result - the *sum rule* of convex analysis. Before we can cite a version of this theorem, we have to introduce the notion of the *core* of a subset M of a Banach space Z (cf. [4, Section 4.1.3]):

$$z \in \mathrm{core}(M) \subset Z \;\Leftrightarrow \bigcup_{\lambda>0} \lambda(M - z) = Z.$$

In particular, $0 \in \mathrm{core}(M)$, if and only if for all $z \in Z$ there are $\lambda > 0$ and $z_m \in M$ such that $z = \lambda z_m$. In this case, M is also called absorbing.

For convex sets M and on Banach spaces, $\mathrm{core}\, M = \mathrm{int}\, M$ by an open-mapping type theorem. It is, however easier to verify $0 \in \mathrm{core}\, M$ than $0 \in \mathrm{int}\, M$.

Theorem B.2 (Sum Rule and Chain Rule)**.** *Let X, Z be Banach spaces and $L : X \to Z$ a continuous linear operator. Let $f : X \to \overline{\mathbb{R}}$ and $g : Z \to \overline{\mathbb{R}}$ be convex and lower semi-continuous functions. If the regularity condition*

$$0 \in \mathrm{core}(\mathrm{dom}\, g - L\,\mathrm{dom}\, f) \tag{73}$$

holds, then

$$\partial(g \circ L + f)(x) = L^*\partial g(Lx) + \partial f(x) \qquad \forall\, x \in X. \tag{74}$$

The condition (73) *is equivalent to the statement: for each $z \in Z$ there is $\lambda > 0$, $z_g \in \mathrm{dom}\, g$, and $x_f \in \mathrm{dom}\, f$, such that $z = \lambda(z_g - Lx_f)$.*

Proof. This variant of the sum-rule can be found, for example, in [4, Theorem 4.3.3]. The regularity condition (73) is classical and goes back to [16] and [15]. □

The reader may recognize similarities with the surjectivity condition for cones imposed in [21]. Actually, this condition could also be used for our theory. However, Theorem B.2 seems to be more natural and convenient for our purpose.