

ICAM

An Optimal Control Problem for Flows with Discontinuities

Eugene M. Cliff

Matthias Heinkenschloss

Ajit R. Shenoy

Interdisciplinary Center for Applied Mathematics
and Department of Mathematics
Virginia Polytechnic Institute and State University

ICAM REPORT 95-09-02

INTERDISCIPLINARY CENTER FOR APPLIED MATHEMATICS
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
BLACKSBURG, VA 24061-0531

SEPTEMBER 1995 (REVISED FEBRUARY 1996)¹

¹This document was generated November 9, 1998

An Optimal Control Problem for Flows with Discontinuities [†]

Eugene M. Cliff Matthias Heinkenschloss

Ajit R. Shenoy

Interdisciplinary Center for Applied Mathematics

Virginia Polytechnic Institute and State University

Blacksburg, Virginia 24061–0531

September 1995 (revised February 1996)

Abstract

In this paper we study a design problem for a duct flow with a shock. The presence of the shock causes numerical difficulties. Good shock-capturing schemes with low continuity properties often cannot be combined successfully with efficient optimization methods requiring smooth functions. A remedy studied in this paper is to introduce the shock location as an explicit variable. This allows one to fit the shock and yields a problem with sufficiently smooth functions. We prove the existence of optimal solutions, Fréchet differentiability, and the existence of Lagrange multipliers. In the second part we introduce and investigate the discrete problem and study the relations between the optimality conditions for the infinite dimensional problem and the discretized one. This reveals information important for the numerical solution of the problem. Numerical examples are given to demonstrate the theoretical findings.

Key words Optimal control, Euler flow equations, sequential quadratic programming.

AMS subject classifications 49M37, 49K15

1 Introduction

The steady flow of an inviscid fluid in a duct of variable cross sectional area $A(x)$ is governed by the Euler equations

$$\mathcal{F}_x + \mathcal{G} = 0, \quad 0 \leq x \leq 1, \quad (1.1)$$

where

$$\mathcal{F} = \begin{pmatrix} (\rho u)A \\ (\rho u^2 + p)A \\ (\rho E + p)uA \end{pmatrix}, \quad \mathcal{G} = \begin{pmatrix} 0 \\ -pA_x \\ 0 \end{pmatrix}.$$

Standard notation has been used, with ρ being the fluid density, u the velocity, $E \equiv e + \dot{u}^2/2$, where e is specific internal energy, and p is the fluid static pressure. The subscript x denotes differentiation with respect

[†]This research was supported by the Air Force Office of Scientific Research under Grants F49620-93-1-0280 and F49620-96-1-0329 and by the NSF under Grant DMS-9403699. A shortened version of this report appeared as [6]. The journal publication [6] should be cited instead of this report.

to the position along the streamwise direction x . Details on the derivation of the Euler equations may be found in any book on compressible fluid dynamics, such as Anderson [3]. It is assumed that the cross-sectional area $A(x)$ of the duct along the streamwise direction x is absolutely continuous and monotonically increasing:

$$A(x) > 0, \quad A_x(x) > 0, \quad x \in [0, 1]. \quad (1.2)$$

Frank and Shubin [8] note that solutions of this simple model exhibit phenomena quite similar to those in two dimensional inviscid flow over an airfoil.

In this paper we are interested in a design problem governed by the equations (1.1). More precisely, given a desired velocity u^d we want to find an area profile $A(x)$ obeying (1.2) and generating a flow that best fits the desired flow u^d in the least squares sense. In a first formulation the design problem can be stated as follows:

$$\text{Minimize } \int_0^1 (u(x) - u^d(x))^2 dx \quad (1.3)$$

over all u and A satisfying (1.1), (1.2), and certain boundary conditions to be specified. A detailed formulation of the state equation will be given in Section 2.

This design problem is difficult to solve numerically, because the duct flow has a shock. Many good numerical shock-capturing schemes, such as the Godunov scheme have low continuity properties, see e.g. [11]. On the other hand, efficient numerical optimization schemes require sufficiently smooth cost and constraint functions. A straightforward combination of off-the-shelf discretization schemes for the flow equations and of off-the-shelf optimization methods often leads to very unsatisfactory results. Frank and Shubin [8] have considered this problem. They have shown that the Euler equations (1.1) can be reduced to a single ODE. For the numerical solution of the design problem they used standard schemes for the discretization of the flow equation and then applied two optimization approaches to solve the discretized problem. The first optimization approach considered in [8] is the black-box method in which the flow variables are considered to be functions of the design variables A , implicitly defined by the flow equation. The resulting optimization problem is then solved using a Newton-like method. The other optimization approach considered was the all-at-once method in which the flow and the design are treated as independent variables. The flow equations are viewed as constraints of the optimization problem. A sequential quadratic programming (SQP) method is applied to the solution of this optimization problem. One of the findings reported in [8] is that their implementations of the black-box approach were more robust than their implementation of the all-at-once method. Even though there may be a lack of smoothness in the functions, the black-box-approach using e.g. finite difference derivative approximations seems to tolerate this. However, if the all-at-once method converged, then, as expected, it was much faster than the black-box approach.

The failure of the SQP method is related to the presence of shocks in the flow. In particular, if shock-capturing schemes with low continuity properties, like the Godunov scheme, are used, then the SQP method, which is designed for smooth problems, behaves poorly. This seeming incompatibility of good shock-capturing schemes for the discretization of the flow equations and efficient SQP methods for the solution of the optimization problem motivated the research in this paper. Our approach to this problem is an extended and refined version of the all-at-once approach in [8]. The important extension is that we include the shock location as a state variable. This guarantees differentiability of the problem. Moreover, we use a discretization of the flow equations that is very similar to the Godunov scheme. The formulation gives a sharp shock and, since the shock location is an explicit variable, the map from the design parameters to the flow is differentiable. In this paper, we give a rigorous analysis of the infinite dimensional design problem including existence of optimal designs, existence of Fréchet derivatives, and existence of Lagrange multipliers. In particular we will show that the co-state is discontinuous at the shock location, unless the

target velocity can be matched perfectly. The second part of the paper is devoted to the numerical solution of the problem. We discuss the discretized design problem and investigate the relation between the finite dimensional problem and the discretized one. The careful study of this relation gives valuable insight and reveals information that is shown to be important for the performance of the optimization algorithm.

Other aspects of this design problem and other methods for its solution have been studied by Borggaard [4], by Shenoy and Cliff [16] and by Narducci, Grossman, and Haftka [15]. In [4] the flow variables are viewed as functions of the designs and a sensitivity equation approach is used to compute the gradient. An optimal control approach is used in [16]. There it is assumed that the shock location is known. The derivative A_x of the area is the control variable and the area and the flow are the state variables. The design problem is then formulated as an optimal control problem governed by a system of ODEs and solved using a multiple shooting method to solve the two-point boundary value problem which is obtained from the necessary optimality conditions. In [15] sensitivities for various (semi-) discretizations are studied. Although the shock location is treated implicitly, it enters the discretization scheme for the objective function. The design problem is solved numerically using the black-box-approach with a steepest descent method.

2 The One-Dimensional Nozzle Flow

It has been shown by Frank and Shubin [8] that equation (1.1) can be reduced to a single ordinary differential equation in u . In fact, (1.1) is equivalent to the ODE

$$(f(u))_x + g(u, A) = 0, \quad (2.1)$$

where

$$f(u) \equiv u + \bar{H}/u, \quad g(u, A) \equiv \frac{A_x}{A}(\bar{\gamma}u - \bar{H}/u), \quad (2.2)$$

and where

$$\bar{\gamma} = (\gamma - 1)/(\gamma + 1), \quad \bar{H} = 2H\bar{\gamma}$$

are given constants. The constant $\gamma > 1$ is the gas constant (for air, $\gamma = 1.4$), and the constant H is the total enthalpy. The flow is supersonic for $u > u_* \equiv \sqrt{\bar{H}}$ and subsonic for $u < u_*$.

In addition, we impose the following boundary conditions

$$u(0) = u_{\text{in}}, \quad u(1) = u_{\text{out}}. \quad (2.3)$$

We choose boundary data $u_{\text{in}} > u_* > u_{\text{out}}$ so that a solution u of (2.1), (2.3) has a jump from supersonic to subsonic at some point x_s . At the shock location x_s the flow is required to satisfy the Rankine–Hugoniot relation

$$f(u(x_s-)) = f(u(x_s+)) \quad (2.4)$$

or, equivalently,

$$u(x_s-) \cdot u(x_s+) = \bar{H}. \quad (2.5)$$

As usual, $u(x_s-) = \lim_{h \rightarrow 0+} u(x_s - h)$ and $u(x_s+) = \lim_{h \rightarrow 0+} u(x_s + h)$. Equation (2.1) along with the above conditions (2.3) and (2.4) defines the flow profile.

For sake of completeness, we review the results presented by Frank and Shubin [8] concerning the existence of solutions to (2.1), (2.4), and (2.3). This existence result will not be needed in this form, but the arguments applied for its proof give some important insight into the structure of the problem.

First we consider the initial value problems

$$(f(u))_x + g(u, A) = 0, \quad u(0) = u_{\text{in}} \quad (2.6)$$

and

$$(f(u))_x + g(u, A) = 0, \quad u(1) = u_{\text{out}}. \quad (2.7)$$

Since $f_u(u) > 0$ for $u > \sqrt{\bar{H}}$, there exists a solution of (2.6) in a neighborhood $[0, x_L)$ of $x = 0$ provided $u_{\text{in}} > \sqrt{\bar{H}}$. If $u_{\text{in}} \in (\sqrt{\bar{H}}, \sqrt{2\bar{H}})$ and $u(1) = u_{\text{out}} < \sqrt{\bar{H}}$, then (1.2) and the definitions of f, g imply that

$$u_x(x) = -g(u(x), A(x))/f_u(u(x)) \begin{cases} > 0 & x = 0, \\ < 0 & x = 1. \end{cases}$$

Using the continuity of the solution and bootstrapping, we can deduce that unique solutions of (2.6) and (2.7) exist on some interval $[0, x_L)$ and $(x_R, 1]$, respectively. Moreover, the solutions are monotonically increasing and decreasing, respectively. It is easy to verify that the solutions are implicitly given by

$$A(x)u(x)(2H - u^2(x))^r = \begin{cases} K_L, & x \in [0, x_L), \\ K_R, & x \in (x_R, 1], \end{cases} \quad (2.8)$$

where $r = 1/(\gamma - 1)$ and where the constants K_L, K_R are determined from $A(0), u(0) = u_{\text{in}}$ and $A(1), u(1) = u_{\text{out}}$. Due to the restrictions on the boundary conditions and due to the fact that $A(x) > 0$, the constants K_L, K_R are positive. Equation (2.8) defines two functions $A_L(u), A_R(u)$. It can be easily verified that these functions have a minimum at $u_* = \sqrt{\bar{H}}$ and are strictly monotone on $(0, \sqrt{\bar{H}}]$ and on $[\sqrt{\bar{H}}, \sqrt{2\bar{H}})$. Thus, the initial condition $u_{\text{in}} \in (\sqrt{\bar{H}}, \sqrt{2\bar{H}})$ guarantees that the solution u of (2.6) exists on $[0, x_L) = [0, \infty)$. The point x_R is the uniquely defined point satisfying $A(x_R) = A_R^*$. The situation is sketched in Figure 2.1.

Using (2.5), (2.8), and the continuity of A , the shock position x_s can be characterized by

$$w(u(x_s-)) = \frac{1}{K_L}u(x_s-)(2H - u^2(x_s-))^r - \frac{1}{K_R}\frac{\bar{H}}{u(x_s-)}\left(2H - \left(\frac{\bar{H}}{u(x_s-)}\right)^2\right)^r = 0. \quad (2.9)$$

It is easy to see that $\lim_{u \rightarrow \sqrt{2\bar{H}}-} w(u) < 0$. Hence, given A there exists a boundary conditions $u_{\text{in}} \in (\sqrt{\bar{H}}, \sqrt{2\bar{H}})$, $u_{\text{out}} \in (0, \sqrt{\bar{H}})$, i.e. K_L, K_R , such that $\lim_{u \rightarrow u_{\text{in}}+} w(u) > 0$. In this case there exists $u(x_s-)$ such that $w(u(x_s-)) = 0$. Since the area A is monotonically increasing, the shock condition can then be computed from (2.8). Thus, we can conclude the following result:

Theorem 2.1 *Suppose the area function satisfies (1.2). Then there exist boundary conditions $u_{\text{in}} \in (\sqrt{\bar{H}}, \sqrt{2\bar{H}})$ and $u_{\text{out}} \in (0, \sqrt{\bar{H}})$ such that the equations (2.1), (2.4), and (2.3) admit a unique solution u which is supersonic and monotonically increasing on $(0, x_s)$ and subsonic and monotonically decreasing on $(x_s, 1)$. Moreover, it obeys the inequalities*

$$\begin{aligned} \sqrt{\bar{H}} < u_{\text{in}} \leq u(x) < \min \left\{ \sqrt{2\bar{H}}, \bar{H}/u_{\text{out}} \right\}, & x \in [0, x_s), \\ u_{\text{out}} \leq u(x) < \bar{H}/u_{\text{in}} < \sqrt{\bar{H}}, & x \in (x_s, 1]. \end{aligned} \quad (2.10)$$

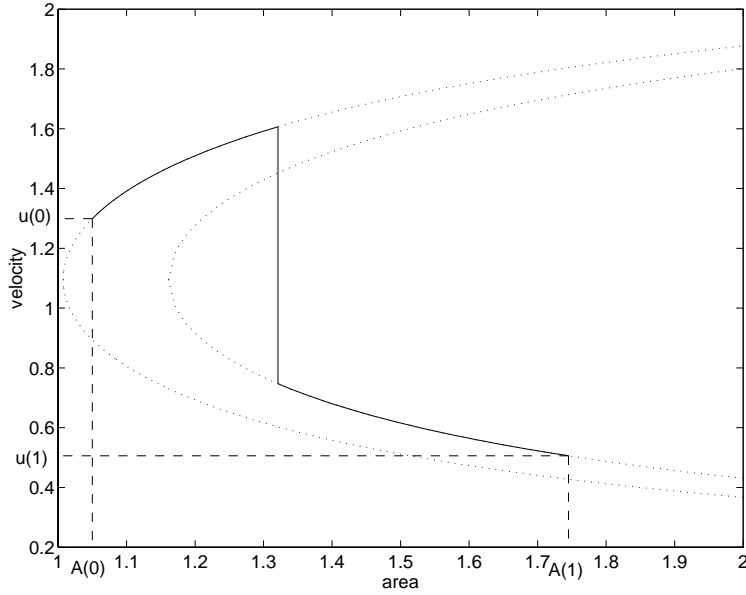


Figure 2.1: Sketch of the velocity as a function of the area.

In the following we will denote the logarithmic derivative of A by q ,

$$q(x) \equiv \frac{A_x(x)}{A(x)} = \frac{d}{dx} \ln(A(x)). \quad (2.11)$$

With this substitution the state equation is given by

$$(f(u))_x + g(u, q) = 0, \quad (2.12)$$

and (2.3), (2.5), where

$$g(u, q) \equiv q(\bar{\gamma}u - \bar{H}/u). \quad (2.13)$$

Note that instead of introducing another symbol we redefine g . Unless stated otherwise, in the following $g(u, q)$ is always given by (2.13).

Trivially, q is determined by A . On the other hand, if the area is known at some point, for example, if $A(0) > 0$ is given, then $A(x)$ can be computed from q by integrating (2.11):

$$A(x) = \exp\left(\ln(A(0)) + \int_0^x q(t) dt\right). \quad (2.14)$$

The function $A(x)$ defined by (2.14) is absolutely continuous and, therefore, differentiable almost everywhere. From now on, we assume that $A(0) > 0$ is given.

For a rigorous treatment of the dependence of the solution upon parameters, we have to transform the ODE, the shock condition, and the boundary conditions. We denote the velocity left of the shock by u_L , the velocity right of the shock by u_R , and, as before, the shock location by x_s . If we perform the variable transformation $x \rightarrow x_s \xi$ left of the shock and $x \rightarrow 1 - (1 - x_s)\xi$ right of the shock and if we set

$$u_L(\xi) = u(x_s \xi), \quad u_R(\xi) = u(1 - (1 - x_s)\xi), \quad (2.15)$$

then we find that

$$\frac{d}{d\xi}u_L(\xi) = x_s \frac{d}{dx}u(x_s\xi), \quad \frac{d}{d\xi}u_R(\xi) = (x_s - 1) \frac{d}{dx}u(1 - (1 - x_s)\xi). \quad (2.16)$$

Thus, in the new spatial variables the ODE (2.12), the shock condition (2.5), and the boundary conditions (2.3) become

$$(f(u_L))_\xi + x_s g(u_L, q_L) = 0, \quad \xi \in [0, 1], \quad (2.17)$$

$$(f(u_R))_\xi + (x_s - 1)g(u_R, q_R) = 0, \quad \xi \in [0, 1], \quad (2.18)$$

$$f(u_L(1)) = f(u_R(1)), \quad (2.19)$$

and

$$u_L(0) = u_{\text{in}}, \quad u_R(0) = u_{\text{out}}. \quad (2.20)$$

The functions q_L and q_R in (2.17), (2.18) are defined by $q_L(\xi) = q(x_s\xi)$ and $q_R(\xi) = q(1 - (1 - x_s)\xi)$, respectively. In the following we view q_L and q_R as independent variables. To guarantee that the corresponding area function is monotonically increasing we have to impose the conditions

$$q_L(\xi) > 0, \quad q_R(\xi) > 0, \quad \xi \in [0, 1].$$

3 The Design Problem

The control problem we are interested in is the design of an area function generating a flow that best approximates a desired velocity in the least squares sense. Given a desired velocity $u^d \in L^2(0, 1)$ and a point x_s we introduce

$$u_L^d(x_s; \xi) = u^d(x_s\xi), \quad u_R^d(x_s; \xi) = u^d(1 - (1 - x_s)\xi). \quad (3.1)$$

With the transformations (2.15), (3.1) the objective function (1.3) is given by

$$\begin{aligned} \int_0^1 (u(x) - u^d(x))^2 dx &= \int_0^{x_s} (u(x) - u^d(x))^2 dx + \int_{x_s}^1 (u(x) - u^d(x))^2 dx \\ &= x_s \int_0^1 (u_L(\xi) - u_L^d(x_s; \xi))^2 d\xi + (1 - x_s) \int_0^1 (u_R(\xi) - u_R^d(x_s; \xi))^2 d\xi. \end{aligned}$$

Thus, using the transformation introduced at the end of Section 2, the control problem we have to solve is given as follows:

$$\min J(u_L, u_R, x_s, q_L, q_R) \equiv \frac{x_s}{2} \int_0^1 (u_L(\xi) - u_L^d(x_s; \xi))^2 d\xi + \frac{1 - x_s}{2} \int_0^1 (u_R(\xi) - u_R^d(x_s; \xi))^2 d\xi \quad (3.2)$$

subject to the equality constraints

$$(f(u_L))_\xi + x_s g(u_L, q_L) = 0, \quad \xi \in [0, 1], \quad (3.3)$$

$$(f(u_R))_\xi + (x_s - 1)g(u_R, q_R) = 0, \quad \xi \in [0, 1], \quad (3.4)$$

$$f(u_L(1)) = f(u_R(1)), \quad (3.5)$$

$$u_L(0) = u_{\text{in}}, \quad u_R(0) = u_{\text{out}}, \quad (3.6)$$

and to the inequality constraints

$$0 \leq x_s \leq 1, \quad (3.7)$$

$$0 \leq q_{\text{low}} \leq q_L(\xi), q_R(\xi) \leq q_{\text{upp}}, \quad \xi \in [0, 1]. \quad (3.8)$$

We assume that the boundary conditions obey

$$u_{\text{in}} \in (\sqrt{H}, \sqrt{2H}), \quad u_{\text{out}} \in (\sqrt{\gamma H}, \sqrt{H}). \quad (3.9)$$

The states are given by the triple (u_L, u_R, x_s) and the controls are (q_L, q_R) . The equations (3.3), (3.4), (3.5), (3.6) are the state equations.

As the control space we use

$$\mathcal{Q} = L^\infty([0, 1]) \times L^\infty([0, 1])$$

and we denote the set of admissible controls by

$$\mathcal{Q}_{ad} = \{(q_L, q_R) \in \mathcal{Q} \mid 0 \leq q_{\text{low}} \leq q_L(\xi), q_R(\xi) \leq q_{\text{upp}} \text{ a.e. in } [0, 1]\}.$$

The set of admissible controls is closed and convex. By a solution to (3.3) (or (3.4)) we mean an absolutely continuous function which satisfies (3.3) (or (3.4)) almost everywhere on $[0, 1]$.

Using the arguments applied in the previous section we can establish the following result.

Lemma 3.1 *Suppose that u_{in} and u_{out} obey (3.9) and that $(q_L, q_R) \in \mathcal{Q}_{ad}$. If (u_L, u_R, x_s) with $x_s \in [0, 1]$ is a solution of (3.3)–(3.6), then*

$$0 < u_{\text{out}} \leq u_R(\xi) \leq \frac{\bar{H}}{u_{\text{in}}} < \sqrt{\bar{H}} < u_{\text{in}} \leq u_L(\xi) \leq \frac{\bar{H}}{u_{\text{out}}} < \sqrt{2\bar{H}}, \quad \xi \in [0, 1], \quad (3.10)$$

and there exists $c > 0$ which depends only on $u_{\text{in}}, u_{\text{out}}$ and $q_{\text{low}}, q_{\text{upp}}$ such that

$$|(u_L)_\xi(\xi)| \leq c, \quad |(u_R)_\xi(\xi)| \leq c \quad \text{a.e. on } [0, 1]. \quad (3.11)$$

Proof: The estimate (3.10) follows from the monotonicity properties of the solution and the Rankine–Hugoniot relation written in the form (2.5). See also the discussion in Section 2.

From (3.3) and (3.4) we find that

$$\begin{aligned} (u_L)_\xi(\xi) &= -\frac{x_s g(u_L(\xi), q_L(\xi))}{f_u(u_L(\xi))} = -\frac{x_s q_L(\xi)(\bar{\gamma}u_L(\xi) - \bar{H}/u_L(\xi))}{1 - \bar{H}/u_L^2(\xi)} \quad \text{a.e. on } [0, 1], \\ (u_R)_\xi(\xi) &= -\frac{(x_s - 1)g(u_R(\xi), q_R(\xi))}{f_u(u_R(\xi))} = -\frac{(x_s - 1)q_R(\xi)(\bar{\gamma}u_R(\xi) - \bar{H}/u_R(\xi))}{1 - \bar{H}/u_R^2(\xi)} \quad \text{a.e. on } [0, 1]. \end{aligned}$$

The function $\bar{H}/u - \bar{\gamma}u = (\bar{\gamma}/u)(2H - u^2)$ is monotonically decreasing in u and positive for $u \in (0, \sqrt{2\bar{H}})$. Using the estimate (3.10) it can be seen that

$$|(u_L)_\xi(\xi)| \leq \frac{q_{\text{upp}}(\bar{H}/u_{\text{in}} - \bar{\gamma}u_{\text{in}})}{1 - \bar{H}/u_{\text{in}}^2}, \quad |(u_R)_\xi(\xi)| \leq \frac{q_{\text{upp}}(\bar{H}/u_{\text{out}} - \bar{\gamma}u_{\text{out}})}{u_{\text{in}}^2/\bar{H} - 1} \quad \text{a.e. on } [0, 1].$$

□

Thus, the state space appropriate for this design problem is given by

$$\mathcal{U} = W^{1,\infty}([0, 1]) \times W^{1,\infty}([0, 1]) \times \mathbb{R}$$

In the following we simply write $W^{1,\infty}$, L^∞ instead of $W^{1,\infty}([0, 1])$, $L^\infty([0, 1])$ and we set

$$\|q\|_\infty = \text{ess sup}_{[0,1]} |q(\xi)|, \quad \|u\|_{1,\infty} = \|u\|_\infty + \|u_\xi\|_\infty.$$

We note that the shock location x_s enters the design problem in the differential equations (3.3), (3.4) and in the objective function, see also (3.1). The functions u_L , u_R , q_L , and q_R do not depend explicitly on x_s , but are implicitly coupled with the shock location through the design problem, in particular through (3.3) and (3.4).

Theorem 3.2 *Suppose there exist $x_s \in (0, 1)$ and $\bar{u} \in (u_{\text{in}}, \bar{H}/u_{\text{out}})$ such that*

$$0 \leq q_{\text{low}} \leq \min \left\{ \frac{\left(1 - \frac{\bar{H}}{u_{\text{in}}^2}\right)(\bar{u} - u_{\text{in}})}{x_s \left(\frac{\bar{H}}{u_{\text{in}}} - \bar{\gamma} u_{\text{in}}\right)}, \frac{\left(\frac{\bar{u}^2}{\bar{H}} - 1\right)\left(\frac{\bar{H}}{\bar{u}} - u_{\text{out}}\right)}{(1 - x_s)\left(\frac{\bar{H}}{u_{\text{out}}} - \bar{\gamma} u_{\text{out}}\right)} \right\} \quad (3.12)$$

and

$$q_{\text{upp}} \geq \max \left\{ \frac{(1 - \bar{\gamma})(\bar{u} - u_{\text{in}})}{x_s \left(\frac{\bar{H}}{\bar{u}} - \bar{\gamma} \bar{u}\right)}, \frac{\left(\frac{\bar{H}}{u_{\text{out}}} - 1\right)\left(\frac{\bar{H}}{\bar{u}} - u_{\text{out}}\right)}{(1 - x_s)\left(\bar{u} - \bar{\gamma} \frac{\bar{H}}{\bar{u}}\right)} \right\}. \quad (3.13)$$

Then there exists an optimal control $(q_L^*, q_R^*) \in \mathcal{Q}$ of (3.2) – (3.8).

Proof: First, note that the conditions (3.9) guarantee that the min in the condition (3.12) is positive.

(i) Existence of feasible points: For given $x_s \in (0, 1)$ and $\bar{u} \in (u_{\text{in}}, \bar{H}/u_{\text{out}})$ we set

$$u_L(\xi) = u_{\text{in}} + \xi(\bar{u} - u_{\text{in}}), \quad u_R(\xi) = u_{\text{out}} + \xi\left(\frac{\bar{H}}{\bar{u}} - u_{\text{out}}\right),$$

and

$$q_L(\xi) = -\frac{\left(1 - \frac{\bar{H}}{u_L^2(\xi)}\right)(\bar{u} - u_{\text{in}})}{x_s \left(\bar{\gamma} u_L(\xi) - \frac{\bar{H}}{u_L(\xi)}\right)}, \quad q_R(\xi) = -\frac{\left(1 - \frac{\bar{H}}{u_R^2(\xi)}\right)\left(\frac{\bar{H}}{\bar{u}} - u_{\text{out}}\right)}{(x_s - 1)\left(\bar{\gamma} u_R(\xi) - \frac{\bar{H}}{u_R(\xi)}\right)}.$$

By construction, $(u_L, u_R, x_s, q_L, q_R)$ satisfies the constraints (3.3)–(3.7).

The function $\bar{H}/u - \bar{\gamma}u = (\bar{\gamma}/u)(2\bar{H} - u^2)$ is monotonically decreasing in u and positive for $u \in (0, \sqrt{2\bar{H}})$. Notice that (3.9) implies the inequalities $\bar{u} < \sqrt{2\bar{H}}$ and $\bar{H}/\bar{u} < \sqrt{2\bar{H}}$. Therefore the functions q_L, q_R obey

$$\frac{\left(1 - \frac{\bar{H}}{u_{\text{in}}^2}\right)(\bar{u} - u_{\text{in}})}{x_s \left(\frac{\bar{H}}{u_{\text{in}}} - \bar{\gamma} u_{\text{in}}\right)} \leq q_L(\xi) \leq \frac{\left(1 - \frac{\bar{H}}{2\bar{H}}\right)(\bar{u} - u_{\text{in}})}{x_s \left(\frac{\bar{H}}{\bar{u}} - \bar{\gamma} \bar{u}\right)} = \frac{(1 - \bar{\gamma})(\bar{u} - \bar{u})}{x_s \left(\frac{\bar{H}}{\bar{u}} - \bar{\gamma} \bar{u}\right)}$$

and

$$\frac{\left(\frac{\bar{u}^2}{\bar{H}} - 1\right)\left(\frac{\bar{H}}{\bar{u}} - u_{\text{out}}\right)}{(1 - x_s)\left(\frac{\bar{H}}{u_{\text{out}}} - \bar{\gamma} u_{\text{out}}\right)} \leq q_R(\xi) \leq \frac{\left(\frac{\bar{H}}{u_{\text{out}}} - 1\right)\left(\frac{\bar{H}}{\bar{u}} - u_{\text{out}}\right)}{(1 - x_s)\left(\bar{u} - \bar{\gamma} \frac{\bar{H}}{\bar{u}}\right)}.$$

Thus, q_L, q_R also satisfy the bound constraints (3.8).

(ii) Existence of optimal controls: This part of the existence result uses standard techniques. Let $\{(u_L^n, u_R^n, x_s^n, q_L^n, q_R^n)\}$ be a minimizing sequence.

By Lemma 3.1 the states obey (3.10) for all n and the derivatives of u_L^n, u_R^n are uniformly bounded. Therefore, the sequence $\{(u_L^n, u_R^n)\}$ is equicontinuous and, by the Arzelà–Ascoli theorem, relatively compact in $C([0, 1])^2$. Thus, there exists a subsequence, for simplicity also denoted $\{n\}$, with

$$\begin{aligned} (u_L^n, u_R^n) &\rightarrow (u_L^*, u_R^*) \quad \text{in } C([0, 1])^2, \\ (q_L^n, q_R^n) &\rightarrow (q_L^*, q_R^*) \quad \text{weak-* in } (L^\infty)^2, \\ x_s^n &\rightarrow x_s^*. \end{aligned}$$

Consider the set $S_k = \{q_L^* \leq q_{\text{low}} - 1/k\}$. Let $m(S_k)$ be the (Lebesgue) measure of this set and let χ_{S_k} be the characteristic function. If $m(S_k) > 0$, then the definition of weak-* convergence implies that

$$m(S_k) q_{\text{low}} \leq \int_0^1 q_L^n(\xi) \chi_{S_k}(\xi) d\xi \rightarrow \int_0^1 q_L^*(\xi) \chi_{S_k}(\xi) d\xi \leq m(S_k) (q_{\text{low}} - 1/k),$$

which is a contradiction. Hence $m(S_k) = 0$ for all k . With $\{q_L^* < q_{\text{low}}\} = \cup_{k \in \mathbb{N}} S_k$ we find that $m(\{q_L^* < q_{\text{low}}\}) = 0$. Using analogous arguments we can deduce that $(q_L^*, q_R^*) \in \mathcal{Q}_{ad}$.

Since u_L^n, u_R^n satisfy (3.5), u_L^*, u_R^* satisfy (3.5). Moreover, using (3.10),

$$\frac{x_s^n (\bar{\gamma} u_L^n - \bar{H}/u_L^n)}{1 - \bar{H}/(u_L^n)^2} \rightarrow \frac{x_s^* (\bar{\gamma} u_L^* - \bar{H}/u_L^*)}{1 - \bar{H}/(u_L^*)^2} \quad \text{in } C([0, 1]).$$

Hence, by taking the limit in

$$u_L^n(\xi) = u_{\text{in}} - \int_0^\xi q_L^n \frac{x_s^n (\bar{\gamma} u_L^n - \bar{H}/u_L^n)}{1 - \bar{H}/(u_L^n)^2} d\zeta$$

we find that

$$u_L^*(\xi) = u_{\text{in}} - \int_0^\xi q_L^* \frac{x_s^* (\bar{\gamma} u_L^* - \bar{H}/u_L^*)}{1 - \bar{H}/(u_L^*)^2} d\zeta,$$

i.e. u_L^* satisfies (3.3). Similarly, we can show that u_R^* satisfies (3.4).

Thus, $(u_L^*, u_R^*, x_s^*, q_L^*, q_R^*)$ is a solution to the optimal control problem. \square

4 Fréchet Differentiability

In the following we view u_L, u_R, x_s and q_L, q_R as independent variables. Since the shock location is treated explicitly, Fréchet differentiability of the objective function and the function of constraints can be established. In this section we introduce the mathematical framework that permits us to prove Fréchet differentiability, derive the first derivatives and we prove the continuous invertibility of the partial Fréchet derivative of the constraints with respect to the state variables. The latter property is important to show that constraint qualifications hold and is essential in SQP methods, in which one has to solve linearized state equations.

We introduce the operator

$$C : \mathcal{U} \times \mathcal{Q} \rightarrow \mathcal{C}, \tag{4.1}$$

where

$$\mathcal{C} = L^\infty \times L^\infty \times \mathbb{R}^3.$$

The operator C is defined as follows: For $r = (r_L, r_R, r_s, r_{\text{in}}, r_{\text{out}}) \in \mathcal{C}$ the equality

$$C(u_L, u_R, x_s, q_L, q_R) = r$$

holds if and only if

$$(f(u_L))_\xi + x_s g(u_L, q_L) = r_L, \quad \xi \in [0, 1], \quad (4.2)$$

$$(f(u_R))_\xi + (x_s - 1)g(u_R, q_R) = r_R, \quad \xi \in [0, 1], \quad (4.3)$$

$$f(u_L(1)) - f(u_R(1)) = r_s, \quad (4.4)$$

and

$$u_L(0) - u_{\text{in}} = r_{\text{in}}, \quad u_R(0) - u_{\text{out}} = r_{\text{out}}. \quad (4.5)$$

The equation $C(u_L, u_R, x_s, q_L, q_R) = 0$ is equivalent to (2.17), (2.18), (2.19), (2.20).

To be able to evaluate (4.2) and (4.3) the velocities have to satisfy $u_L(x) \neq 0$, $u_R(x) \neq 0$ for all $x \in [0, 1]$.

Theorem 4.1 *The nonlinear operator $C : \mathcal{U} \times \mathcal{Q} \rightarrow \mathcal{C}$ is Fréchet differentiable at any point $(u_L, u_R, x_s, q_L, q_R) \in \mathcal{U} \times \mathcal{Q}$ satisfying $u_L(x) \neq 0$, $u_R(x) \neq 0$ for all $x \in [0, 1]$. The partial Fréchet derivatives are given by*

$$C_{(u_L, u_R, x_s)}(u_L, u_R, x_s, q_L, q_R) (\hat{u}_L, \hat{u}_R, \hat{x}_s) = \begin{pmatrix} (f_u(u_L)\hat{u}_L)_\xi + x_s g_u(u_L, q_L)\hat{u}_L + \hat{x}_s g(u_L, q_L) \\ (f_u(u_R)\hat{u}_R)_\xi + (x_s - 1)g_u(u_R, q_R)\hat{u}_R + \hat{x}_s g(u_R, q_R) \\ f_u(u_L(1))\hat{u}_L(1) - f_u(u_R(1))\hat{u}_R(1) \\ \hat{u}_L(0) \\ \hat{u}_R(0) \end{pmatrix}$$

and

$$C_{(q_L, q_R)}(u_L, u_R, x_s, q_L, q_R) (\hat{q}_L, \hat{q}_R) = \begin{pmatrix} x_s g_q(u_L, q_L)\hat{q}_L \\ (x_s - 1)g_q(u_R, q_R)\hat{q}_R \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Proof: To prove the differentiability of C with respect to (u_L, u_R, x_s) we have to show that

$$\begin{aligned} & \left\| C(u_L + \hat{u}_L, u_R + \hat{u}_R, x_s + \hat{x}_s, q_L, q_R) - C(u_L, u_R, x_s, q_L, q_R) \right. \\ & \quad \left. - C_{(u_L, u_R, x_s)}(u_L, u_R, x_s, q_L, q_R)(\hat{u}_L, \hat{u}_R, \hat{x}_s) \right\|_{\mathcal{C}} = o\left(\|(\hat{u}_L, \hat{u}_R, \hat{x}_s)\|_{\mathcal{U}}\right). \end{aligned}$$

For the first component corresponding to the equation (4.2) we obtain

$$\begin{aligned} & (f(u_L + \hat{u}_L))_\xi + (x_s + \hat{x}_s)g(u_L + \hat{u}_L, q_L) - (f(u_L))_\xi - x_s g(u_L, q_L) \\ & - (f_u(u_L)\hat{u}_L)_\xi - x_s g_u(u_L, q_L)\hat{u}_L - \hat{x}_s g(u_L, q_L) \\ & = (f(u_L + \hat{u}_L) - f(u_L) - f_u(u_L)\hat{u}_L)_\xi + x_s [g(u_L + \hat{u}_L, q_L) - g(u_L, q_L) - g_u(u_L, q_L)\hat{u}_L] \end{aligned}$$

$$\begin{aligned}
& +\hat{x}_s [g(u_L + \hat{u}_L, q_L) - g(u_L, q_L)] \\
= & \left(\int_0^1 (f_u(u_L + t\hat{u}_L) - f_u(u_L)) \hat{u}_L dt \right)_\xi + x_s o(\|\hat{u}_L\|_\infty) + \hat{x}_s O(\|\hat{u}_L\|_\infty) \\
= & \int_0^1 (f_{uu}(u_L + t\hat{u}_L)(u_L + t\hat{u}_L)_\xi - f_{uu}(u_L)(u_L)_\xi) \hat{u}_L + (f_u(u_L + t\hat{u}_L) - f_u(u_L))(\hat{u}_L)_\xi dt \\
& + x_s o(\|\hat{u}_L\|_\infty) + \hat{x}_s O(\|\hat{u}_L\|_\infty) \\
= & O(\|\hat{u}_L\|_{1,\infty}^2) + x_s o(\|\hat{u}_L\|_\infty) + \hat{x}_s O(\|\hat{u}_L\|_\infty).
\end{aligned}$$

Similar estimates can be applied to show analogous results for the equations (4.3) and (4.4). This proves the differentiability of C with respect to (u_L, u_R, x_s) .

The differentiability of C with respect to (q_L, q_R) follows easily, since the function g in (2.13) is linear in q . \square

The following result concerns the invertibility of the partial Fréchet derivative $C'_{(u_L, u_R, x_s)}(u_L, u_R, x_s, q_L, q_R)$. For given $(u_L, u_R, x_s) \in \mathcal{U}$, $(q_L, q_R) \in \mathcal{Q}$, and $(r_L, r_R, r_s, r_{in}, r_{out}) \in \mathcal{C}$ we consider the system

$$(f_u(u_L)\hat{u}_L)_\xi + x_s g_u(u_L, q_L)\hat{u}_L + \hat{x}_s g(u_L, q_L) = r_L, \quad \xi \in [0, 1], \quad (4.6)$$

$$(f_u(u_R)\hat{u}_R)_\xi + (x_s - 1)g_u(u_R, q_R)\hat{u}_R + \hat{x}_s g(u_R, q_R) = r_R, \quad \xi \in [0, 1], \quad (4.7)$$

$$f_u(u_L(1))\hat{u}_L(1) - f_u(u_R(1))\hat{u}_R(1) = r_s, \quad (4.8)$$

and

$$\hat{u}_L(0) = r_{in}, \quad \hat{u}_R(0) = r_{out}. \quad (4.9)$$

Theorem 4.2 (i) Suppose that $(u_L, u_R, x_s, q_L, q_R) \in \mathcal{U} \times \mathcal{Q}_{ad}$ is a point satisfying

$$u_i(\xi) \neq 0, \quad u_i(\xi) \neq \sqrt{H}, \quad \forall \xi \in [0, 1], \quad i = L, R, .$$

If

$$\begin{aligned}
& \int_0^1 \exp\left(-\int_t^1 \frac{x_s g_u(u_L(s), q_L(s))}{f_u(u_L(s))} ds\right) g(u_L(t), q_L(t)) dt \\
\neq & \int_0^1 \exp\left(-\int_t^1 \frac{(x_s - 1)g_u(u_R(s), q_R(s))}{f_u(u_R(s))} ds\right) g(u_R(t), q_R(t)) dt, \quad (4.10)
\end{aligned}$$

then for every $(r_L, r_R, r_s, r_{in}, r_{out}) \in \mathcal{C}$ the system (4.6), (4.7), (4.8), (4.9) admits a unique solution $(\hat{u}_L, \hat{u}_R, \hat{x}_s) \in \mathcal{U}$ which depends continuously on $(r_L, r_R, r_s, r_{in}, r_{out})$.

(ii) If, in addition, there exists constants δ, Δ with $0 < \delta < \Delta$ such that the point $(u_L, u_R, x_s, q_L, q_R)$ obeys

$$\delta \leq u_i(\xi) \leq \Delta, \quad |u_i(\xi) - \sqrt{H}| \geq \delta, \quad \forall \xi \in [0, 1], \quad i = L, R,$$

and

$$\left| \int_0^1 \exp\left(-\int_t^1 \frac{x_s g_u(u_L, q_L)}{f_u(u_L)} ds\right) g(u_L, q_L) - \exp\left(-\int_t^1 \frac{(x_s - 1)g_u(u_R, q_R)}{f_u(u_R)} ds\right) g(u_R, q_R) dt \right| > \delta,$$

then there exists a constant K dependent on δ, Δ , but independent of $(u_L, u_R, x_s, q_L, q_R)$ such that

$$\|(\hat{u}_L, \hat{u}_R, \hat{x}_s)\|_{\mathcal{U}} \leq K \|(r_L, r_R, r_s, r_{in}, r_{out})\|_{\mathcal{C}}.$$

Proof: (i) First we note that since $W^{1,\infty}(0,1) \subset C([0,1])$, there exists $\delta > 0$ such that $u_i(\xi) > \delta$, $|u_i(\xi) - \sqrt{H}| > \delta$, for all $\xi \in [0,1]$, $i = L, R$.

The equation (4.6) is equivalent to

$$(f_u(u_L)\hat{u}_L)_\xi + \frac{x_s g_u(u_L, q_L)}{f_u(u_L)} (f_u(u_L)\hat{u}_L) = r_L - \hat{x}_s g(u_L, q_L). \quad (4.11)$$

Using the integrating factor

$$\mu_L(\xi) = \exp\left(\int_0^\xi \frac{x_s g_u(u_L(t), q_L(t))}{f_u(u_L(t))} dt\right),$$

the solution of (4.11) with initial condition $\hat{u}_L(0) = r_{\text{in}}$ is given by

$$\hat{u}_L(\xi) = + \frac{1}{\mu_L(\xi) f_u(u_L(\xi))} \left(r_{\text{in}} f_u(u_L(0)) + \int_0^\xi \mu_L(t) [r_L(t) - \hat{x}_s g(u_L(t), q_L(t))] dt \right). \quad (4.12)$$

Similarly, one can show that the solution of (4.7) with initial condition $\hat{u}_R(0) = r_{\text{out}}$ is given by

$$\hat{u}_R(\xi) = \frac{1}{\mu_R(\xi) f_u(u_R(\xi))} \left(r_{\text{out}} f_u(u_R(0)) + \int_0^\xi \mu_R(t) [r_R(t) - \hat{x}_s g(u_R(t), q_R(t))] dt \right), \quad (4.13)$$

where

$$\mu_R(\xi) = \exp\left(\int_0^\xi \frac{(x_s - 1) g_u(u_R(t), q_R(t))}{f_u(u_R(t))} dt\right).$$

Inserting (4.12), (4.13) into (4.8) yields

$$\begin{aligned} & r_{\text{in}} \frac{f_u(u_L(0))}{\mu_L(1)} + \int_0^1 \exp\left(-\int_t^1 \frac{x_s g_u(u_L(s), q_L(s))}{f_u(u_L(s))} ds\right) [r_L(t) - \hat{x}_s g(u_L(t), q_L(t))] dt \\ = & r_{\text{out}} \frac{f_u(u_R(0))}{\mu_R(1)} + \int_0^1 \exp\left(-\int_t^1 \frac{(x_s - 1) g_u(u_R(s), q_R(s))}{f_u(u_R(s))} ds\right) [r_R(t) - \hat{x}_s g(u_R(t), q_R(t))] dt \end{aligned} \quad (4.14)$$

If the inequality (4.10) is valid, then (4.14) can be solved for \hat{x}_s . This proves the existence and uniqueness of the solution.

The continuous dependence of $(\hat{u}_L, \hat{u}_R, \hat{x}_s) \in \mathcal{U}$ upon $(r_L, r_R, r_s, r_{\text{in}}, r_{\text{out}}) \in \mathcal{C}$ follows from the equations (4.12), (4.13), (4.14).

(ii) The assertion follows from equations (4.12), (4.13), and (4.14). \square

Corollary 4.3 *If $(\bar{u}_L, \bar{u}_R, \bar{x}_s, \bar{q}_L, \bar{q}_R) \in \mathcal{U} \times \mathcal{Q}$ is feasible, i.e. it satisfies the constraints (3.3)–(3.8), and if there exists $\delta > 0$ with*

$$\left| \int_0^1 \exp\left(-\int_t^1 \frac{x_s g_u(\bar{u}_L, \bar{q}_L)}{f_u(\bar{u}_L)} ds\right) g(\bar{u}_L, \bar{q}_L) - \exp\left(-\int_t^1 \frac{(x_s - 1) g_u(\bar{u}_R, \bar{q}_R)}{f_u(\bar{u}_R)} ds\right) g(\bar{u}_R, \bar{q}_R) dt \right| > \delta,$$

then there exists $\epsilon > 0$ such that for all $(u_L, u_R, x_s) \in \mathcal{U}$, $(q_L, q_R) \in \mathcal{Q}$ with

$$\|(u_L, u_R, x_s) - (\bar{u}_L, \bar{u}_R, \bar{x}_s)\|_{\mathcal{U}} < \epsilon, \quad \|(q_L, q_R) - (\bar{q}_L, \bar{q}_R)\|_{\mathcal{Q}} < \epsilon$$

and for all $(r_L, r_R, r_s, r_{\text{in}}, r_{\text{out}}) \in \mathcal{C}$ the system (4.6), (4.7), (4.8), (4.9) admits a unique solution $(\hat{u}_L, \hat{u}_R, \hat{x}_s) \in \mathcal{U}$. Moreover, there exists a constant K independent of $(\bar{u}_L, \bar{u}_R, \bar{x}_s, \bar{q}_L, \bar{q}_R)$ such that

$$\|(\hat{u}_L, \hat{u}_R, \hat{x}_s)\|_{\mathcal{U}} \leq K \|(r_L, r_R, r_s, r_{\text{in}}, r_{\text{out}})\|_{\mathcal{C}}.$$

Proof: The solutions u_L, u_R satisfy (3.10). Hence, the assertion follows from Theorem 4.2(ii). \square

From the definitions of f and g one can see that the Fréchet derivative of C is Lipschitz–continuous for all u_L, u_R with $u_L(\xi), u_R(\xi) \geq \underline{u} > 0$ for all $\xi \in [0, 1]$. Moreover, C is even twice Fréchet differentiable if $u_L, u_R > 0$.

To prove the Fréchet differentiability of the objective function we have to keep in mind that the desired velocity depends on x_s , cf. (3.1). Therefore differentiability with respect to x_s can only be guaranteed if the desired velocity u^d is sufficiently smooth, a fact that will be addressed again in the numerical examples section.

Theorem 4.4 *If the desired velocity u^d is differentiable with absolutely continuous derivative, then the objective function J is Fréchet differentiable. The partial Fréchet –derivatives are given by*

$$\begin{aligned} J_{u_L}(u_L, u_R, x_s, q_L, q_R) \hat{u}_L &= x_s \int_0^1 (u_L(\xi) - u_L^d(x_s; \xi)) \hat{u}_L(\xi) d\xi, \\ J_{u_R}(u_L, u_R, x_s, q_L, q_R) \hat{u}_R &= (1 - x_s) \int_0^1 (u_R(\xi) - u_R^d(x_s; \xi)) \hat{u}_R(\xi) d\xi, \end{aligned}$$

and

$$\begin{aligned} &J_{x_s}(u_L, u_R, x_s, q_L, q_R) \\ &= \int_0^1 \frac{1}{2} (u_L(\xi) - u_L^d(x_s; \xi))^2 - x_s (u_L(\xi) - u_L^d(x_s; \xi)) (u_L^d)_x(\xi) \xi d\xi \\ &\quad - \int_0^1 \frac{1}{2} (u_R(\xi) - u_R^d(x_s; \xi))^2 + (1 - x_s) (u_R(\xi) - u_R^d(x_s; \xi)) (u_R^d)_x(\xi) \xi d\xi. \end{aligned} \quad (4.15)$$

The objective function J is twice Fréchet differentiable if the desired velocity u^d is twice differentiable with absolutely continuous second derivative.

Proof: The assertion follows from the definition of J using standard estimates. The proof is therefore omitted. \square

We conclude this section with a brief discussion of the differentiability of the velocity function. Suppose that we have an area function \bar{q}_L, \bar{q}_R and corresponding velocities \bar{u}_L, \bar{u}_R and shock location \bar{x}_s that satisfy the state equations (3.3), (3.4), (3.5), (3.6) and are such that (4.10) is fulfilled. Then the implicit function theorem guarantees the differentiability of the function

$$L^\infty \times L^\infty \ni (q_L, q_R) \longrightarrow (u_L, u_R, x_s) \in W^{1,\infty} \times W^{1,\infty} \times \mathbb{R}$$

that maps the area into the solution of the state equation at this point. In fact, the derivative is given by

$$(u_L(\bar{q}_L, \bar{q}_R), u_L(\bar{q}_L, \bar{q}_R), x_s(\bar{q}_L, \bar{q}_R))_{(\bar{q}_L, \bar{q}_R)} = -C_{(u_L, u_R, x_s)}(\bar{u}_L, \bar{u}_R, \bar{x}_s, \bar{q}_L, \bar{q}_R)^{-1} C_{(q_L, q_R)}(\bar{u}_L, \bar{u}_R, \bar{x}_s, \bar{q}_L, \bar{q}_R).$$

Given u_L, u_R, x_s , the velocity of the original problem can be obtained via (2.15), i.e.

$$u(x) = \begin{cases} u_L\left(\frac{x}{x_s}\right), & x \in [0, x_s], \\ u_R\left(\frac{1-x}{1-x_s}\right), & x \in [x_s, 1]. \end{cases} \quad (4.16)$$

If one considers the map

$$W^{1,\infty} \times W^{1,\infty} \times \mathbb{R} \ni (u_L, u_R, x_s) \longrightarrow u \in L^\infty$$

that is defined by (4.16), then it is easy to see that because of the presence of a shock this map is not Fréchet differentiable. In fact it is not even continuous. This shows that differentiability is only lost when the composite function

$$(q_L, q_R) \longrightarrow (u_L, u_R, x_s) \longrightarrow u$$

is considered. If left and right velocity and shock location are treated as independent variables, then, as shown in this section, differentiability can be guaranteed under suitable assumptions.

5 Optimality Conditions

We define the Lagrange function

$$\begin{aligned} & L(u_L, u_R, x_s, q_L, q_R, \lambda_L, \lambda_R, \lambda_s) \\ &= \frac{x_s}{2} \int_0^1 (u_L - u_L^d)^2 d\xi + \frac{1-x_s}{2} \int_0^1 (u_R - u_R^d)^2 d\xi + \int_0^1 \lambda_L [(f(u_L))_\xi + x_s g(u_L, q_L)] d\xi \\ &+ \int_0^1 \lambda_R [(f(u_R))_\xi + (x_s - 1)g(u_R, q_R)] d\xi + \lambda_s [f(u_L(1)) - f(u_R(1))]. \end{aligned} \quad (5.1)$$

If the shock location at the optimum obeys $x_s \in (0, 1)$, then the first order necessary optimality conditions are

$$\begin{aligned} 0 &= L_{(u_L, u_R, x_s)}(u_L, u_R, x_s, q_L, q_R, \lambda_L, \lambda_R, \lambda_s)(\hat{u}_L, \hat{u}_R, \hat{x}_s), \\ 0 &\leq L_{(q_L, q_R)}(u_L, u_R, x_s, q_L, q_R, \lambda_L, \lambda_R, \lambda_s)(\hat{q}_L, \hat{q}_R), \\ 0 &= L_{(\lambda_L, \lambda_R, \lambda_s)}(u_L, u_R, x_s, q_L, q_R, \lambda_L, \lambda_R, \lambda_s)(\hat{\lambda}_L, \hat{\lambda}_R, \hat{\lambda}_s), \end{aligned} \quad (5.2)$$

for all $(\hat{u}_L, \hat{u}_R, \hat{x}_s)$ with $\hat{u}_L(0) = \hat{u}_R(0) = 0$, for all (\hat{q}_L, \hat{q}_R) with $(q_L + \hat{q}_L, q_R + \hat{q}_R) \in \mathcal{Q}_{ad}$, and for all $(\hat{\lambda}_L, \hat{\lambda}_R, \hat{\lambda}_s)$.

The third equation in (5.2) yields the state equation (3.3), (3.4), (3.5), (3.6). Using integration by parts we find that the first equation in (5.2) with $\hat{u}_L(0) = \hat{u}_R(0) = 0$ yields

$$\begin{aligned} 0 &= L_{(u_L, u_R, x_s)}(u_L, u_R, x_s, q_L, q_R, \lambda_L, \lambda_R, \lambda_s)(\hat{u}_L, \hat{u}_R, \hat{x}_s) \\ &= x_s \int_0^1 (u_L - u_L^d) \hat{u}_L d\xi + (1-x_s) \int_0^1 (u_R - u_R^d) \hat{u}_R d\xi + J_{x_s}(u_L, u_R, x_s, q_L, q_R) \hat{x}_s \\ &+ \int_0^1 -(\lambda_L)_\xi f_u(u_L) \hat{u}_L + \lambda_L [x_s g_u(u_L, q_L) \hat{u}_L + \hat{x}_s g(u_L, q_L)] d\xi \\ &+ \int_0^1 -(\lambda_R)_\xi f_u(u_R) \hat{u}_R + \lambda_R [(x_s - 1)g_u(u_R, q_R) \hat{u}_R + \hat{x}_s g(u_R, q_R)] d\xi \\ &+ (\lambda_L(1) + \lambda_s) f_u(u_L(1)) \hat{u}_L(1) + (\lambda_R(1) - \lambda_s) f_u(u_R(1)) \hat{u}_R(1). \end{aligned} \quad (5.3)$$

If one sets $\hat{x}_s = 0$ and varies over all (\hat{u}_L, \hat{u}_R) with $\hat{u}_L(0) = \hat{u}_R(0) = \hat{u}_L(1) = \hat{u}_R(1) = 0$, then one obtains the adjoint equations

$$(\lambda_L)_\xi f_u(u_L) = x_s g_u(u_L, q_L) \lambda_L + x_s (u_L - u_L^d), \quad (5.4)$$

$$(\lambda_R)_\xi f_u(u_R) = (x_s - 1) g_u(u_R, q_R) \lambda_R + (1 - x_s)(u_R - u_R^d). \quad (5.5)$$

Allowing $\hat{u}_L(1), \hat{u}_R(1) \neq 0$ yields the conditions

$$\lambda_L(1) = -\lambda_s, \quad \lambda_R(1) = \lambda_s. \quad (5.6)$$

Finally, varying \hat{x}_s gives

$$\int_0^1 \lambda_L g(u_L, q_L) + \lambda_R g(u_R, q_R) d\xi + J_{x_s}(u_L, u_R, x_s, q_L, q_R) = 0. \quad (5.7)$$

The existence of Lagrange multipliers are guaranteed if the operator of linearized constraints is onto. Thus, existence of Lagrange multipliers is expected under the assumptions of Theorem 4.2(i). We provide a proof of this result, since the explicit form of the Lagrange multipliers derived in the proof are of interest in connection with the discretized problem.

Theorem 5.1 *If the assumptions of Theorem 4.2(i) are valid, then the adjoint system (5.4), (5.5), (5.6), (5.7) admits a unique solution.*

Proof: Equation (5.4) is equivalent to

$$(\lambda_L)_\xi - \frac{x_s g_u(u_L, q_L)}{f_u(u_L)} \lambda_L = x_s \frac{u_L - u_L^d}{f_u(u_L)}.$$

Using the integrating factor

$$\nu_L(\xi) = \exp \left(\int_\xi^1 \frac{x_s g_u(u_L(t), q_L(t))}{f_u(u_L(t))} dt \right),$$

the solution of (5.4) with $\lambda_L(1) = -\lambda_s$ is given by

$$(\lambda_L)(\xi) = \frac{1}{\nu_L(\xi)} \left(-\lambda_s - \int_\xi^1 x_s \nu_L \frac{u_L - u_L^d}{f_u(u_L)} dt \right). \quad (5.8)$$

Similarly, the solution of (5.5) with $\lambda_R(1) = \lambda_s$ is given by

$$(\lambda_R)(\xi) = \frac{1}{\nu_R(\xi)} \left(\lambda_s - \int_\xi^1 (1 - x_s) \nu_R \frac{u_R - u_R^d}{f_u(u_R)} dt \right), \quad (5.9)$$

where

$$\nu_R(\xi) = \exp \left(\int_\xi^1 \frac{(x_s - 1) g_u(u_R(t), q_R(t))}{f_u(u_R(t))} dt \right).$$

Inserting the solutions into (5.7), we find that

$$\begin{aligned}
& \int_0^1 \exp\left(-\int_\xi^1 \frac{x_s g_u(u_L, q_L)}{f_u(u_L)} ds\right) g(u_L, q_L) - \exp\left(-\int_\xi^1 \frac{(x_s - 1)g_u(u_R, q_R)}{f_u(u_R)} ds\right) g(u_R, q_R) d\xi \lambda_s \\
= & J_{x_s}(u_L, u_R, x_s, q_L, q_R) - \int_0^1 \left\{ \int_\xi^1 (1 - x_s) \exp\left(-\int_\xi^t \frac{(x_s - 1)g_u(u_R, q_R)}{f_u(u_R)} ds\right) \frac{u_R - u_R^d}{f_u(u_R)} dt \right. \\
& \left. + \int_\xi^1 x_s \exp\left(-\int_\xi^t \frac{x_s g_u(u_L, q_L)}{f_u(u_L)} ds\right) \frac{u_L - u_L^d}{f_u(u_L)} dt \right\} d\xi. \tag{5.10}
\end{aligned}$$

Since (4.10) holds true, equation (5.10) has a unique solution λ . This concludes the proof of the theorem. \square

The second equation in (5.2) is equivalent to

$$\lambda_L(\xi) x_s \left(\bar{\gamma} u_L(\xi) - \frac{\bar{H}}{u_L(\xi)} \right) \begin{cases} \geq 0 & \text{if } q_L(\xi) = q_{\text{low}}, \\ = 0 & \text{if } q_L(\xi) \in (q_{\text{low}}, q_{\text{upp}}), \\ \leq 0 & \text{if } q_L(\xi) = q_{\text{upp}}, \end{cases} \tag{5.11}$$

and

$$\lambda_R(\xi) (x_s - 1) \left(\bar{\gamma} u_R(\xi) - \frac{\bar{H}}{u_R(\xi)} \right) \begin{cases} \geq 0 & \text{if } q_R(\xi) = q_{\text{low}}, \\ = 0 & \text{if } q_R(\xi) \in (q_{\text{low}}, q_{\text{upp}}), \\ \leq 0 & \text{if } q_R(\xi) = q_{\text{upp}}. \end{cases} \tag{5.12}$$

We conclude this section with a brief discussion of the continuity properties of the Lagrange multiplier. The co-states λ_L and λ_R obey the shock condition (5.6). From the examination of the other equations it can be seen that $\lambda_L(1) = -\lambda_R(1) = 0$ can be guaranteed only if the desired velocities can be matched exactly, i.e. the source terms in (5.4), (5.5), and the term $\mathcal{J}_{x_s}(u_L, u_R, x_s, q_L, q_R)$ in (5.7) vanish. For a related design problem governed by the full Euler equations Iollo et. al. [13] concluded that the co-state is continuous and zero at the shock location. Even though Iollo et. al. [13] do not use a shock fitting method and even though the problem investigated in this paper is a reduction of the one-dimensional Euler equations and therefore a simplification of the problem studied in [13] our result indicates a different behavior of the co-states. The fact that for this problem the co-states are nonzero at the shock if the optimal value of the objective function is nonzero can also be observed in the numerical experiments. See Section 7. We have, in fact, generalized the results [5] obtained in this section to the full 1-D Euler equations (1.1), and our analysis does indicate nonzero and discontinuous behavior of the co-state at the shock location. Our findings have not been corroborated with numerical results, as of yet.

6 The Discrete Design Problem

As in the analysis of the continuous problem we divide the interval into two subintervals $[0, x_s]$ and $[x_s, 1]$. The shock location x_s is one of the state variables. The transformation onto the fixed domain as shown at the end of Section 2, however, is not performed explicitly, but incorporated implicitly using moving grids on the left and on the right of the shock. As we will see later, this is equivalent to discretizing the fixed domain control problem (3.2) – (3.8).

For the discretization of the optimal control problem we use a cell centered grid. The subinterval $[0, x_s]$ left of the shock is subdivided into N_L equidistant subintervals of length $h_L = x_s/N_L$, the subinterval

$[x_s, 1]$ right of the shock is subdivided into N_R equidistant subintervals of length $h_R = (1 - x_s)/N_R$. The point x_i denotes the midpoint of the i th cell:

$$\begin{aligned} x_i &= (i - \frac{1}{2})h_L, & i &= 1, \dots, N_L, \\ x_i &= x_s + (i - \frac{1}{2} - N_L)h_R, & i &= N_L + 1, \dots, N_L + N_R. \end{aligned} \quad (6.1)$$

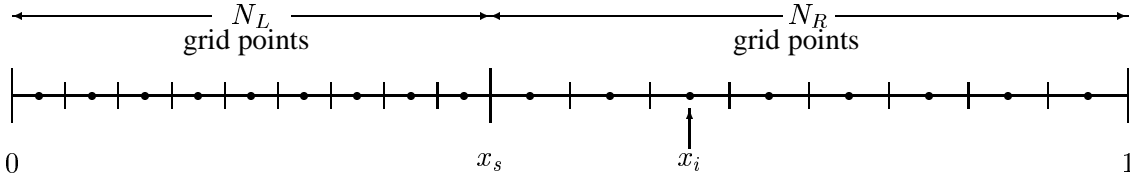


Figure 6.1: The Grid

On each cell the function q defining the area of the duct and the velocity u are approximated by constants q_i and u_i , respectively. Therefore the number of state variables is $N_L + N_R + 1$ and the number of control variables is $N_L + N_R$.

The objective function is discretized using the midpoint rule:

$$\frac{x_s}{2} \int_0^1 (u_L(\xi) - u_L^d(x_s; \xi))^2 d\xi + \frac{1 - x_s}{2} \int_0^1 (u_R(\xi) - u_R^d(x_s; \xi))^2 d\xi \approx J^h(u, x_s, q),$$

where

$$J^h(u, x_s, q) \equiv \frac{1}{2} \sum_{i=1}^{N_L} h_L (u_i - u^d(x_i))^2 + \frac{1}{2} \sum_{i=N_L+1}^{N_L+N_R} h_R (u_i - u^d(x_i))^2. \quad (6.2)$$

For the discretization of the differential equation (2.12) we use

$$\frac{f_{i+1/2} - f_{i-1/2}}{\Delta x_i} + g(u_i, q_i) = 0,$$

where Δx_i denotes the width of cell i . The fluxes at the cell boundaries are approximated as follows: In the supersonic region left of the shock we set $f_{i+1/2} = f(u_i)$ giving

$$\frac{f(u_i) - f(u_{i-1})}{h_L} + g(u_i, q_i) = 0, \quad i = 1, \dots, N_L. \quad (6.3)$$

In the subsonic region right of the shock we set $f_{i+1/2} = f(u_{i+1})$ giving

$$\frac{f(u_{i+1}) - f(u_i)}{h_R} + g(u_i, q_i) = 0, \quad i = N_L + 1, \dots, N_L + N_R. \quad (6.4)$$

The Rankine–Hugoniot condition (2.4) is discretized as

$$f(u_{N_L}) - f(u_{N_L+1}) = 0. \quad (6.5)$$

The equations (6.3), (6.4) are also the ones used in the Godunov scheme for the supersonic and subsonic regions, respectively. See e.g. [8].

If one multiplies (6.3) by x_s and (6.4) by $(x_s - 1)$, then one can see that the resulting equations are implicit discretization schemes for (3.3) and (3.4). For (3.3) the indices $i = 0$ and $i = N_L$ correspond to the boundaries $\xi = 0$ and $\xi = 1$, respectively, whereas for (3.4) the indices $i = N_L + N_R$ and $i = N_L + 1$ correspond to the boundaries $\xi = 0$ and $\xi = 1$, respectively.

The equations (6.3) and (6.4) multiplied by x_s and $x_s - 1$, respectively, and the equation (6.5) form the $N_L + N_R + 1$ state constraints

$$C_i^h(u, x_s, q) = 0, \quad i = 1, \dots, N_L + N_R + 1,$$

where

$$C_i^h(u, x_s, q) \equiv \begin{cases} N_L (f(u_i) - f(u_{i-1})) + x_s g(u_i, q_i) & i = 1, \dots, N_L, \\ N_R (-f(u_{i+1}) + f(u_i)) + (x_s - 1) g(u_i, q_i) & i = N_L + 1, \dots, N_L + N_R, \\ f(u_{N_L}) - f(u_{N_L+1}) & i = N_L + N_R + 1. \end{cases} \quad (6.6)$$

The scalars u_0 and $u_{N_L+N_R+1}$ are determined from the boundary conditions (3.6).

This leads to the finite dimensional optimal control problem

$$\begin{aligned} \min \quad & J^h(u, x_s, q) \\ \text{s.t.} \quad & C^h(u, x_s, q) = 0, \\ & 0 \leq q_{\text{low}} \leq q \leq q_{\text{upp}}. \end{aligned} \quad (6.7)$$

The state variables in the discrete problem are $(u_1, \dots, u_{N_L+N_R}, x_s)$ and the control variables are $(q_1, \dots, q_{N_L+N_R})$. One may add a state constraint $0 \leq x_s \leq 1$ to (6.7). However, in our numerical experiments the shock was always in the interior.

Under the assumptions of Theorem 4.4 the discretized objective function J^h is differentiable. In fact, due to the discretization, one can even relax the differentiability assumptions on d . The objective function J^h is differentiable if u^d is differentiable at x_i , $i = 1, \dots, N_L + N_R$. In particular, it holds that

$$\begin{aligned} J_{x_s}^h(u, x_s, q) &= \sum_{i=1}^{N_L} \frac{1}{2N_L} (u_i - u^d(x_i))^2 - h_L \frac{i - \frac{1}{2}}{N_L} (u_i - u^d(x_i)) u_x^d(x_i) \\ &\quad - \sum_{i=N_L+1}^{N_L+N_R} \frac{1}{2N_R} (u_i - u^d(x_i))^2 + h_R \left(1 - \frac{i - \frac{1}{2} - N_L}{N_R}\right) (u_i - u^d(x_i)) u_x^d(x_i). \end{aligned}$$

From the definition of f and g , it is easy to see that C^h is differentiable for all u, x_s, q with $u > 0$. The partial Jacobian $C_{(u, x_s)}^h$ of C^h is a bordered matrix given by

$$C_{(u, x_s)}^h(u, x_s, q) = \begin{pmatrix} B_L & 0 & e_L \\ 0 & B_R & e_R \\ d_L^T & d_R^T & 0 \end{pmatrix}, \quad (6.8)$$

where $B_L \in \mathbb{R}^{N_L \times N_L}$ is a lower bidiagonal matrix, $B_R \in \mathbb{R}^{N_R \times N_R}$ is an upper bidiagonal matrix, and $e_L, d_L \in \mathbb{R}^{N_L}$, $e_R, d_R \in \mathbb{R}^{N_R}$. The structure of the matrix reflects the left hand side of the system (4.6),

(4.7), (4.8), (4.9). The partial Jacobian C_q^h of C^h is a $(N_L + N_R + 1) \times (N_L + N_R)$ ‘diagonal’ matrix with diagonal entries given by

$$\left(C_q^h(u, x_s, q) \right)_{ii} = \begin{cases} x_s g_q(u_i, q_i) & i = 1, \dots, N_L, \\ (x_s - 1) g_q(u_i, q_i) & i = N_L + 1, \dots, N_L + N_R. \end{cases}$$

The function g depends linearly on q , c.f. (2.13).

If $C_{(u, x_s)}^h(u, x_s, q)$, B_L, B_R are nonsingular, the linear system

$$\begin{pmatrix} B_L & 0 & e_L \\ 0 & B_R & e_R \\ d_L^T & d_R^T & 0 \end{pmatrix} \begin{pmatrix} \hat{u}_L \\ \hat{u}_R \\ \hat{x}_s \end{pmatrix} = \begin{pmatrix} r_L \\ r_R \\ r_s \end{pmatrix}$$

can be solved using

$$\hat{u}_L = B_L^{-1}(r_L - e_L \hat{x}_s), \quad (6.9)$$

$$\hat{u}_R = B_R^{-1}(r_R - e_R \hat{x}_s), \quad (6.10)$$

where

$$\hat{x}_s = (d_L^T B_L^{-1} e_L + d_R^T B_R^{-1} e_R)^{-1} (d_L^T B_L^{-1} r_L + d_R^T B_R^{-1} r_R - r_s). \quad (6.11)$$

This solution procedure is the discrete version of the procedure applied in the proof of Theorem 4.2 to establish the existence of a unique solution of (4.6), (4.7), (4.8), (4.9). This is also the solution procedure that is used in our numerical examples.

Theorem 6.1 *If*

$$u_i > \sqrt{H}, \quad i = 1, \dots, N_L, \quad (6.12)$$

then B_L is nonsingular. If

$$u_i \in (0, \sqrt{H}), \quad i = N_L + 1, \dots, N_L + N_R, \quad (6.13)$$

then B_R is nonsingular, and if (6.12), (6.13), and

$$\begin{aligned} & \frac{1}{N_L} \sum_{j=1}^{N_L} \left(\prod_{k=j}^{N_L} \frac{N_L f_u(u_k) + x_s g_u(u_k, q_k)}{N_L f_u(u_k)} \right)^{-1} g(u_j, q_j) \\ & \neq \frac{1}{N_R} \sum_{j=N_L+1}^{N_L+N_R} \left(\prod_{k=N_L+1}^j \frac{N_R f_u(u_k) + (x_s - 1) g_u(u_k, q_k)}{N_R f_u(u_k)} \right)^{-1} g(u_j, q_j), \end{aligned} \quad (6.14)$$

then the matrix $C_{(u, x_s)}^h(u, x_s, q)$ is nonsingular.

Proof: The i th equation of the system $B_L \hat{u}_L = r_L - e_L \hat{x}_s$ is given by

$$\left(N_L f_u(u_i) + x_s g_u(u_i, q_i) \right) (\hat{u}_L)_i - N_L f_u(u_{i-1}) (\hat{u}_L)_{i-1} = -g(u_i, q_i) \hat{x}_s + (r_L)_i, \quad i = 1, \dots, N_L, \quad (6.15)$$

where $(\hat{u}_L)_0 = 0$. Condition (6.12) guarantees that $N_L f_u(u_i) + x_s g_u(u_i, q_i) > 0$, $i = 1, \dots, N_L$. With

$$z_i = \frac{N_L f_u(u_{i-1})}{N_L f_u(u_i) + x_s g_u(u_i, q_i)}, \quad w_i = \frac{g(u_i, q_i)}{N_L f_u(u_i) + x_s g_u(u_i, q_i)}, \quad v_i = \frac{(r_L)_i}{N_L f_u(u_i) + x_s g_u(u_i, q_i)},$$

these equation can be written as

$$(\hat{u}_L)_i - z_i(\hat{u}_L)_{i-1} = -w_i \hat{x}_s + v_i, \quad i = 1, \dots, N_L,$$

If we multiply the last difference equation by $1/Z_i$, where $Z_i = \prod_{j=1}^i z_j$, $i = 1, \dots, N_L$, $Z_0 = 1$, then we obtain

$$\frac{1}{Z_i}(\hat{u}_L)_i - \frac{1}{Z_{i-1}}(\hat{u}_L)_{i-1} = -\frac{w_i}{Z_i} \hat{x}_s + \frac{v_i}{Z_i}, \quad i = 1, \dots, N_L,$$

The solution of this difference equation is given by

$$(\hat{u}_L)_i = Z_i \left(\sum_{j=1}^i -\frac{w_j}{Z_j} \hat{x}_s + \frac{v_j}{Z_j} \right), \quad i = 1, \dots, N_L.$$

In particular, it holds that

$$\begin{aligned} (\hat{u}_L)_{N_L} &= \frac{\prod_{k=1}^{N_L-1} N_L f_u(u_k)}{\prod_{k=1}^{N_L} N_L f_u(u_k) + x_s g_u(u_k, q_k)} \times \\ &\left(-\sum_{j=1}^{N_L} \prod_{k=1}^{j-1} \frac{N_L f_u(u_k) + x_s g_u(u_k, q_k)}{N_L f_u(u_k)} g(u_j, q_j) \hat{x}_s + \sum_{j=1}^{N_L} \frac{v_j}{Z_j} \right). \end{aligned} \quad (6.16)$$

The i th equation of the system $B_R \hat{u}_R = r_R - e_R \hat{x}_s$ is given by

$$\left(N_R f_u(u_i) + (x_s - 1) g_u(u_i, q_i) \right) (\hat{u}_R)_i - N_R f_u(u_{i+1}) (\hat{u}_R)_{i+1} = -g(u_i, q_i) \hat{x}_s + (r_R)_i, \quad (6.17)$$

$i = N_L + 1, \dots, N_L + N_R$, where $(\hat{u}_R)_{N_L + N_R + 1} = 0$. As before we can rewrite these equations in the form

$$(\hat{u}_R)_i - z_i(\hat{u}_R)_{i+1} = -w_i \hat{x}_s + v_i, \quad i = N_L + 1, \dots, N_L + N_R,$$

where

$$\begin{aligned} z_i &= \frac{N_R f_u(u_{i+1})}{N_R f_u(u_i) + (x_s - 1) g_u(u_i, q_i)}, \quad w_i = \frac{g(u_i, q_i)}{N_R f_u(u_i) + (x_s - 1) g_u(u_i, q_i)}, \\ v_i &= \frac{(r_R)_i}{N_R f_u(u_i) + (x_s - 1) g_u(u_i, q_i)}. \end{aligned}$$

Note that condition (6.13) implies $N_R f_u(u_i) + (x_s - 1) g_u(u_i, q_i) < 0$, $i = N_L + 1, \dots, N_L + N_R$. If we multiply the last difference equation by $1/Z_i$, where $Z_i = \prod_{j=i}^{N_L + N_R} z_j$, $i = N_L + 1, \dots, N_L + N_R$, $Z_{N_L + N_R + 1} = 1$, then we obtain

$$\frac{1}{Z_i}(\hat{u}_R)_i - \frac{1}{Z_{i+1}}(\hat{u}_R)_{i+1} = -\frac{w_i}{Z_i} \hat{x}_s + \frac{v_i}{Z_i}, \quad i = N_L + 1, \dots, N_L + N_R,$$

The solution of (6.17) is given by

$$(\hat{u}_R)_i = Z_i \left(\sum_{j=i}^{N_L + N_R} -\frac{w_j}{Z_j} \hat{x}_s + \frac{v_j}{Z_j} \right), \quad i = N_L + 1, \dots, N_L + N_R. \quad (6.18)$$

In particular, it holds that

$$(\hat{u}_R)_{N_L+1} = \frac{\prod_{k=N_L+2}^{N_L+N_R} N_R f_u(u_k)}{\prod_{k=N_L+1}^{N_L+N_R} N_R f_u(u_k) + (x_s-1) g_u(u_k, q_k)} \times \left(- \sum_{j=N_L+1}^{N_L+N_R} \prod_{k=j+1}^{N_L+N_R} \frac{N_R f_u(u_k) + (x_s-1) g_u(u_k, q_k)}{N_R f_u(u_k)} g(u_j, q_j) \hat{x}_s + \sum_{j=N_L+1}^{N_L+N_R} \frac{v_j}{Z_j} \right).$$

The last equation $d_L^T \hat{u}_L + d_R^T \hat{u}_R = r_s$ of the system is equivalent to

$$f_u(u_{N_L})(\hat{u}_L)_{N_L} - f_u(u_{N_L+1})(\hat{u}_R)_{N_L+1} = r_s. \quad (6.19)$$

Using the expressions (6.16), (6.18) one can see that the equation (6.19) admits a unique solution \hat{x} if and only if

$$\begin{aligned} & \frac{1}{N_L} \prod_{k=1}^{N_L} \frac{N_L f_u(u_k)}{N_L f_u(u_k) + x_s g_u(u_k, q_k)} \sum_{j=1}^{N_L} \prod_{k=1}^{j-1} \frac{N_L f_u(u_k) + x_s g_u(u_k, q_k)}{N_L f_u(u_k)} g(u_j, q_j) \\ & \neq \frac{1}{N_R} \prod_{k=N_L+1}^{N_L+N_R} \frac{N_R f_u(u_k)}{N_R f_u(u_k) + (x_s-1) g_u(u_k, q_k)} \sum_{j=N_L+1}^{N_L+N_R} \prod_{k=j+1}^{N_L+N_R} \frac{N_R f_u(u_k) + (x_s-1) g_u(u_k, q_k)}{N_R f_u(u_k)} g(u_j, q_j). \end{aligned}$$

This condition is equivalent to (6.14). \square

Remark 6.2 Equation (6.14) is the discretized version of (4.10) with $\bar{e} \approx 1 + x$ and

$$\begin{aligned} \int_{x_j - \frac{1}{2}h_L}^1 \frac{x_s g_u(u_L(s), q_L(s))}{f_u(u_L(s))} ds & \approx \sum_{k=j}^{N_L} \frac{x_s g_u(u_k, q_k)}{N_L f_u(u_k)}, \\ \int_{x_j + \frac{1}{2}h_R}^1 \frac{(x_s - 1) g_u(u_R(s), q_R(s))}{f_u(u_R(s))} ds & \approx \sum_{k=N_L+1}^j \frac{(x_s - 1) g_u(u_k, q_k)}{N_R f_u(u_k)}. \end{aligned}$$

The Lagrange function of the discretized design problem (6.7) is given by

$$\begin{aligned} L(u, x_s, q, \lambda) & = \frac{1}{2} \sum_{i=1}^{N_L} h_L (u_i - u^d(x_i))^2 + \frac{1}{2} \sum_{i=N_L+1}^{N_L+N_R} h_R (u_i - u^d(x_i))^2 \\ & + \sum_{i=1}^{N_L} \lambda_i \left(N_L (f(u_i) - f(u_{i-1})) + x_s g(u_i, q_i) \right) \\ & + \sum_{i=N_L+1}^{N_L+N_R} \lambda_i \left(N_R (-f(u_{i+1}) + f(u_i)) + (x_s - 1) g(u_i, q_i) \right) \\ & + \lambda_{N_L+N_R+1} \left(f(u_{N_L}) - f(u_{N_L+1}) \right). \end{aligned} \quad (6.20)$$

The equations $L_{(u, x_s)}(u, x_s, q, \lambda) = 0$ are equivalent to

$$\begin{aligned} \left(N_L f_u(u_i) + x_s g_u(u_i, q_i) \right) \lambda_i - N_L f_u(u_i) \lambda_{i+1} & = -h_L (u_i - u^d(x_i)), \\ & i = 1, \dots, N_L - 1, \end{aligned} \quad (6.21)$$

$$\left(N_L f_u(u_i) + x_s g_u(u_i, q_i) \right) \lambda_i + f_u(u_i) \lambda_{N_L+N_R+1} = -h_L(u_i - u^d(x_i)), \quad i = N_L, \quad (6.22)$$

$$\left(N_R f_u(u_i) + (x_s - 1) g_u(u_i, q_i) \right) \lambda_i - f_u(u_i) \lambda_{N_L+N_R+1} = -h_R(u_i - u^d(x_i)), \quad i = N_L + 1, \quad (6.23)$$

$$\left(N_R f_u(u_i) + (x_s - 1) g_u(u_i, q_i) \right) \lambda_i - N_R f_u(u_i) \lambda_{i-1} = -h_R(u_i - u^d(x_i)), \quad (6.24)$$

$$i = N_L + 2, \dots, N_L + N_R,$$

and

$$J_{x_s}^h(u, x_s, q) + \sum_{i=1}^{N_L} \lambda_i g(u_i, q_i) + \sum_{i=N_L+1}^{N_L+N_R} \lambda_i g(u_i, q_i) = 0. \quad (6.25)$$

The system (6.21)–(6.25) is given by

$$\begin{pmatrix} B_L^T & 0 & d_L \\ 0 & B_R^T & d_R \\ e_L^T & e_R^T & 0 \end{pmatrix} \begin{pmatrix} \lambda_L \\ \lambda_R \\ \lambda_s \end{pmatrix} = \begin{pmatrix} r_L \\ r_R \\ r_s \end{pmatrix}, \quad (6.26)$$

where we used the notation

$$\lambda_L = (\lambda_1, \dots, \lambda_{N_L}), \quad \lambda_R = (\lambda_{N_L+1}, \dots, \lambda_{N_L+N_R}), \quad \lambda_s = \lambda_{N_L+N_R+1},$$

and

$$r_L = \left(-h_L(u_1 - u^d(x_1)), \dots, -h_L(u_{N_L+1} - u^d(x_{N_L+1})) \right),$$

$$r_R = \left(-h_R(u_{N_L+1} - u^d(x_{N_L+1})), \dots, -h_R(u_{N_L+N_R} - u^d(x_{N_L+N_R})) \right),$$

$$r_s = -J_{x_s}^h(u, x_s, q).$$

The system (6.21)–(6.25) are the adjoint equations of the discretized problem (6.7). However, the equations (6.21)–(6.25) are *not* consistent with the adjoint equations (5.4), (5.5), (5.7).

The inconsistency of the adjoint equations (6.21) to (6.25) of the discretized problem can be removed if we define

$$\begin{aligned} \tilde{\lambda}_i &= N_L \lambda_i, \quad i = 1, \dots, N_L, \\ \tilde{\lambda}_i &= N_R \lambda_i, \quad i = N_L + 1, \dots, N_L + N_R, \\ \tilde{\lambda}_i &= \lambda_i, \quad i = N_L + N_R + 1. \end{aligned} \quad (6.27)$$

In the scaled Lagrange multipliers, the equations (6.21)–(6.24) are equivalent to

$$-\frac{\tilde{\lambda}_{i+1} - \tilde{\lambda}_i}{1/N_L} f_u(u_i) + x_s g_u(u_i, q_i) \tilde{\lambda}_i = -x_s (u_i - u^d(x_i)), \quad (6.28)$$

$$i = 1, \dots, N_L - 1,$$

$$\left(N_L f_u(u_i) + x_s g_u(u_i, q_i) \right) \tilde{\lambda}_i + N_L f_u(u_i) \tilde{\lambda}_{N_L+N_R+1} = -x_s (u_i - u^d(x_i)), \quad i = N_L, \quad (6.29)$$

$$\left(N_R f_u(u_i) + (x_s - 1) g_u(u_i, q_i) \right) \tilde{\lambda}_i - N_R f_u(u_i) \tilde{\lambda}_{N_L+N_R+1} = -(1 - x_s) (u_i - u^d(x_i)), \quad (6.30)$$

$$i = N_L + 1,$$

$$-\frac{\tilde{\lambda}_{i-1} - \tilde{\lambda}_i}{1/N_R} f_u(u_i) + (x_s - 1) g_u(u_i, q_i) \tilde{\lambda}_i = -(1 - x_s) (u_i - u^d(x_i)), \quad (6.31)$$

$$i = N_L + 2, \dots, N_L + N_R.$$

The equations (6.28) and (6.31) are consistent with the infinite dimensional adjoint equations (5.4) and (5.5). Equation (6.28) is an implicit scheme for (5.4) starting at $x = x_s$ and marching towards $x = 0$, the equation (6.31) is an implicit scheme for (5.5) starting at $x = x_s$ and marching towards $x = 1$.

The equations (6.29) and (6.30) are equivalent to

$$f_u(u_i)\tilde{\lambda}_i + f_u(u_i)\tilde{\lambda}_{N_L+N_R+1} + \frac{x_s}{N_L} g_u(u_i, q_i)\tilde{\lambda}_i = -h_L(u_i - u^d(x_i)), \quad i = N_L, \quad (6.32)$$

$$f_u(u_i)\tilde{\lambda}_i - f_u(u_i)\tilde{\lambda}_{N_L+N_R+1} + \frac{x_s-1}{N_R} g_u(u_i, q_i)\tilde{\lambda}_i = -h_R(u_i - u^d(x_i)), \quad i = N_L + 1. \quad (6.33)$$

These equations are consistent with the initial conditions (5.6).

In the scaled Lagrange multipliers, equation (6.25) is written as

$$\sum_{i=1}^{N_L} \tilde{\lambda}_i \frac{1}{N_L} g(u_i, q_i) + \sum_{i=N_L+1}^{N_L+N_R} \tilde{\lambda}_i \frac{1}{N_R} g(u_i, q_i) = -J_{x_s}^h(u, x_s, q) \quad (6.34)$$

which correspond to the equation (5.7).

The system (6.28)–(6.31) and (6.34) is given by

$$\begin{pmatrix} B_L^T & 0 & \tilde{d}_L \\ 0 & B_R^T & \tilde{d}_R \\ \tilde{e}_L^T & \tilde{e}_R^T & 0 \end{pmatrix} \begin{pmatrix} \tilde{\lambda}_L \\ \tilde{\lambda}_R \\ \tilde{\lambda}_s \end{pmatrix} = \begin{pmatrix} \tilde{r}_L \\ \tilde{r}_R \\ \tilde{r}_s \end{pmatrix}, \quad (6.35)$$

where we used the notation

$$\tilde{\lambda}_L = (\tilde{\lambda}_1, \dots, \tilde{\lambda}_{N_L}), \quad \tilde{\lambda}_R = (\tilde{\lambda}_{N_L+1}, \dots, \tilde{\lambda}_{N_L+N_R}), \quad \tilde{\lambda}_s = \tilde{\lambda}_{N_L+N_R+1},$$

and

$$\begin{aligned} \tilde{r}_L &= \left(-x_s(u_1 - u^d(x_1)), \dots, -x_s(u_{N_L} - u^d(x_{N_L})) \right), \\ \tilde{r}_R &= \left(-(1-x_s)(u_{N_L+1} - u^d(x_{N_L+1})), \dots, -(1-x_s)(u_{N_L+N_R} - u^d(x_{N_L+N_R})) \right), \\ \tilde{r}_s &= -J_{x_s}^h(u, x_s, q). \end{aligned}$$

The entries in the system matrices in (6.26) and (6.35) are related as follows:

$$\tilde{d}_L = N_L d_L, \quad \tilde{d}_R = N_R d_R, \quad \tilde{e}_L = \frac{1}{N_L} e_L, \quad \tilde{e}_R = \frac{1}{N_R} e_R.$$

Notice that

$$\begin{pmatrix} B_L & 0 & \tilde{e}_L \\ 0 & B_R & \tilde{e}_R \\ \tilde{d}_L^T & \tilde{d}_R^T & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{N_L} I & 0 & 0 \\ 0 & \frac{1}{N_R} I & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} B_L & 0 & e_L \\ 0 & B_R & e_R \\ d_L^T & d_R^T & 0 \end{pmatrix} \begin{pmatrix} N_L I & 0 & 0 \\ 0 & N_R I & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (6.36)$$

In particular, this relation shows that the system (6.35) is uniquely solvable if and only if (6.26) is uniquely solvable.

The previous considerations raise the question of the “correct” Lagrange multipliers. The Lagrange multipliers λ appear to be the correct ones if one starts with the discrete system. On the other hand, the

Lagrange multipliers $\tilde{\lambda}$ appear to be the appropriate ones if one tries to establish a relation with the original, infinite dimensional problem. This discrepancy can be overcome, if one chooses the appropriate scalar product for the control space.

If we consider u, x_s as a function of the area q defined by the discrete state equation $C^h(u, x_s, q) = 0$, then we can write the discrete optimal control problem (6.7) in the reduced form

$$\begin{aligned} \min \quad & \hat{J}^h(q) \equiv J^h(u(q), x_s(q), q) \\ \text{s.t.} \quad & 0 \leq q_{\text{low}} \leq q \leq q_{\text{upp}}. \end{aligned} \quad (6.37)$$

For the sake of presentation, we assume that for all q with $0 \leq q_{\text{ow}} \leq q \leq q_{\text{upp}}$ the equation $C^h(u, x_s, q) = 0$ has a unique solution. Using the implicit function theorem, the derivative of the reduced objective can be shown to be

$$\hat{J}_q^h(q) \delta q = \left(\nabla_q J^h(u(q), x_s(q), q) + C_q^h(u(q), x_s(q), q) \lambda \right)^T \delta q.$$

See for example [8], [9]. Hence, the gradient of the reduced objective function with respect to the Euclidean scalar product is given by

$$\nabla_q \hat{J}^h(q) = \nabla_q J^h(u(q), x_s(q), q) + C_q^h(u(q), x_s(q), q) \lambda.$$

If we define the scalar product in the discretized control space to be the weighted Euclidean scalar product

$$\langle q_1, q_2 \rangle_{\mathcal{Q}_h} = \sum_{i=1}^{N_L} \frac{1}{N_L} (q_1)_i (q_2)_i + \sum_{i=N_L+1}^{N_L+N_R} \frac{1}{N_R} (q_1)_i (q_2)_i, \quad (6.38)$$

then we find that

$$\hat{J}_q^h(q) \delta q = \left((C_q^h(u, x_s, q))^T \lambda \right)^T \delta q = \langle (C_q^h(u, x_s, q))^T \tilde{\lambda}, \delta q \rangle_{\mathcal{Q}_h}.$$

Thus, the gradient of the reduced objective function with respect to the weighted Euclidean scalar product is given by

$$\nabla_q \hat{J}^h(q) = \nabla_q J^h(u(q), x_s(q), q) + C_q^h(u(q), x_s(q), q) \tilde{\lambda}.$$

Moreover, with (6.36) it is easy to see that the system matrix in (6.35) is the adjoint of the Jacobian $C_{(u, x_s)}^h(u, x_s, q)$ with respect to a weighted scalar product. In fact, if we define

$$\langle \lambda_1, \lambda_2 \rangle_{\Lambda_h} = \sum_{i=1}^{N_L} \frac{1}{N_L} (\lambda_1)_i (\lambda_2)_i + \sum_{i=N_L+1}^{N_L+N_R} \frac{1}{N_R} (\lambda_1)_i (\lambda_2)_i + (\lambda_1)_{N_L+N_R+1} (\lambda_2)_{N_L+N_R+1}, \quad (6.39)$$

then

$$\begin{aligned} & \left\langle \begin{pmatrix} \tilde{\lambda}_L \\ \tilde{\lambda}_R \\ \tilde{\lambda}_s \end{pmatrix}, \begin{pmatrix} B_L & 0 & e_L \\ 0 & B_R & e_R \\ d_L^T & d_R^T & 0 \end{pmatrix} \begin{pmatrix} \tilde{u}_L \\ \tilde{u}_R \\ \tilde{u}_s \end{pmatrix} \right\rangle_{\Lambda_h} \\ &= \begin{pmatrix} \tilde{\lambda}_L \\ \tilde{\lambda}_R \\ \tilde{\lambda}_s \end{pmatrix}^T \begin{pmatrix} \frac{1}{N_L} I & 0 & 0 \\ 0 & \frac{1}{N_R} I & 0 \\ 0^T & 0^T & 1 \end{pmatrix} \begin{pmatrix} B_L & 0 & e_L \\ 0 & B_R & e_R \\ d_L^T & d_R^T & 0 \end{pmatrix} \begin{pmatrix} N_L I & 0 & 0 \\ 0 & N_R I & 0 \\ 0^T & 0^T & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{N_L} I & 0 & 0 \\ 0 & \frac{1}{N_R} I & 0 \\ 0^T & 0^T & 1 \end{pmatrix} \begin{pmatrix} \tilde{u}_L \\ \tilde{u}_R \\ \tilde{u}_s \end{pmatrix} \\ &= \left\langle \begin{pmatrix} B_L^T & 0 & \tilde{d}_L \\ 0 & B_R^T & \tilde{d}_R \\ \tilde{e}_L^T & \tilde{e}_R^T & 0 \end{pmatrix} \begin{pmatrix} \tilde{\lambda}_L \\ \tilde{\lambda}_R \\ \tilde{\lambda}_s \end{pmatrix}, \begin{pmatrix} \tilde{u}_L \\ \tilde{u}_R \\ \tilde{u}_s \end{pmatrix} \right\rangle_{\Lambda_h}. \end{aligned}$$

Note that $\langle q_1, q_2 \rangle_{\mathcal{Q}_h}$ and $\langle \lambda_1, \lambda_2 \rangle_{\Lambda_h}$ are the appropriate discretizations of the scalar products on $L^2(0, 1) \times L^2(0, 1)$ and $L^2(0, 1) \times L^2(0, 1) \times \mathbb{R}$, respectively.

7 Numerical Results

In our numerical experiments the discrete optimal control problem (6.7) is solved using a sequential quadratic programming (SQP) method. These methods solve the nonlinear constrained problem (3.2)–(3.8) or the corresponding discretized problem (6.7) by solving a sequence of quadratic programming problems. We use one of the SQP methods that have been derived and analyzed in [7] and [10]. The SQP methods described there include a trust region globalization to guarantee convergence of the iteration from arbitrary starting points and to enhance its robustness. Moreover, these methods use an affine scaling interior point strategy to handle the bounds on q . The SQP methods incorporate the general structure of optimal control problems like (6.7) and allow for inexact solutions of the quadratic subproblems. In particular, structure in the linearized state equation can be easily incorporated. Moreover the implementation of these SQP algorithms allows the use of weighted scalar products. We apply a reduced SQP method which uses limited memory BFGS updates for the reduced Hessian. The initial Hessian was chosen to be the identity and the number of updates stored will be denoted by L . For more details on this class of SQP methods and their convergence properties we refer to [7] and [10]. In all computations, the trust region was active in the first few iterations. We also point out that since we do not add a regularization term like $\rho(\int_0^1 q_L^2 + \int_0^1 q_R^2)$ in the objective function, the reduced Hessian for the infinite dimensional problem can only be expected to be positive semidefinite. The discretization sometimes has a regularizing effect and in this case for a fixed discretization the reduced Hessian for the discretized problem may be positive definite. However, in this case the smallest eigenvalue converges towards zero as the discretization is refined. The lack of positive definiteness will of course affect the convergence behavior. We have made runs with a regularization term as shown above added to the objective function. As expected, the SQP algorithm required fewer iterations. However, to be compatible with the computations reported in Frank and Shubin [8], we have omitted the regularization term here.

One issue which will be emphasized in this section is the influence of the relation between the infinite dimensional problem and its discretization onto the performance of the optimization method. For the algorithms studied in [7] and [10] no convergence theory in infinite dimensional spaces is known yet. However, convergence results for SQP methods in infinite dimensional spaces are given e.g. in [1], [2], where Lagrange–Newton–SQP methods in Banach spaces using exact second order derivative information are investigated. A convergence analysis of SQP methods in Hilbert spaces allowing quasi–Newton approximations for the second derivative is given in [14]. Here, equality constrained problems are considered, but bound constraints are not included. In all these references, the convergence theory is local and it is assumed that the quadratic programming subproblems are solved exactly. Another reference that is important in this context is [12]. In these references the appropriate implementation of the optimization algorithms for the discretized problems is shown to be important. The underlying infinite dimensional problem dominates the discretized problems. If the discretized problems are treated as finite dimensional nonlinear programming problems, i.e., if the underlying infinite dimensional problem structure is ignored, the performance of the optimization algorithms usually deteriorates as the discretization is refined. The use of weighted scalar products that are obtained from the discretization of the proper scalar products of the infinite dimensional problem emphasize the underlying infinite dimensional character of these problems. The implementation of the SQP algorithm with weighted scalar products as discussed in the previous section is the proper application of the frameworks used in the previous references. In our numerical experiments reported below this leads to a substantial improvement in the performance of the algorithms. We also point out that differen-

tiability of the functions and continuous invertibility of the linearized constraints are important conditions that have to hold in order to formulate the SQP method and to prove its local convergence in the neighborhood of strict local minimizers. For the infinite and finite dimensional version of the design problem, these properties have been established in this paper.

In all our numerical computations we use the constants

$$\gamma = 1.4, \quad H = 3.6, \quad \text{which yield} \quad \bar{\gamma} = 1/6, \quad \bar{H} = 1.2.$$

The target velocity u^d is computed as follows: Given a cubic area function A uniquely determined by

$$A(0) = 1.05, \quad A_x(0) = 0.1, \quad A(1) = 1.745, \quad A_x(1) = 0.1,$$

we use a discretization scheme similar to the one described in this paper to compute the corresponding velocity u as a solution of (2.1), (2.3), and (2.4) with inflow and outflow velocities given by

$$u_{\text{in}} = 1.299, \quad u_{\text{out}} = 0.506.$$

The discretization scheme applied to solve (2.1), (2.3), (2.4) also treats the shock location as an explicit variable and approximates the ODE using a scheme corresponding to (6.6). We use 200 subintervals left of the shock and 200 subintervals right of the shock to compute the target velocity. Note that for the construction of the target velocity the area A and not its logarithmic derivative is used. For the formulation of our optimal design problem we need continuous data. These are obtained by using spline interpolations. First we compute two cubic splines using the points $(x_1, u_1), \dots, (x_{200}, u_{200})$ and $(x_{201}, u_{201}), \dots, (x_{400}, u_{400})$ and then we join these two cubic splines by constructing a cubic polynomial interpolating $(x_{200}, u_{200}), (x_{201}, u_{201})$ and the derivatives of the two previously constructed splines at x_{200} and x_{201} , respectively. The so computed resulting target velocity u^d is continuously differentiable. Note that the area function A and the boundary values $u_{\text{in}}, u_{\text{out}}$ used in the computation of the target data are identical to the ones given in [8]. Unless stated otherwise, we use the bound constraints $q_{\text{ow}} = 0, q_{\text{upp}} = 1$.

The starting values for the SQP method are as follows: The initial logarithmic derivative q of the area is chosen to be $q = 0.5$. The initial shock location is computed from the target data and is chosen to be $x_s = \frac{1}{2}(x_{200} + x_{201})$. For the initial velocity we use a piecewise linear function. On the left of the shock the initial velocity is a linear interpolation between u_n at $x = 0$ and $u_{x_s} = 1.7$ at the initial estimate x_s of the shock location. On the right of the shock we use the linear interpolation $u_{x_s} = \bar{H}/1.7$ and u_{out} . This interpolation scheme guarantees that $u \in (\sqrt{\bar{H}}, \sqrt{2\bar{H}})$ left of the shock and $u < \sqrt{\bar{H}}$ right of the shock. If we would simply use $u_i = u^d(x_i)$, then these restrictions on u would not be satisfied if the initial shock location does not match the target shock location.

With these starting values and target data, and the discretization $N_L = N_R = 100$ the initial function value is $J^h(u, x_s, q) \approx 0.9 * 10^{-3}$ and the norm of the residual is $\|C^h(u, x_s, q)\|_{\Lambda_h} \approx 0.13$. Here, the residual is computed using the weighted norm induced by (6.39). The bound constraints were never active. The necessary optimality conditions show that in this case the Lagrange multipliers $\lambda_1, \dots, \lambda_{N_L+N_R}$ at the optimum are zero. See also (5.11), (5.12). If the truncation criteria $\|C_q^T \lambda\|_{\mathbb{R}^{N_L+N_R}} < \epsilon$ is used, then the Lagrange multipliers $\lambda_1, \dots, \lambda_{N_L+N_R}$ are of the order ϵ . Analogous statements hold for the Lagrange multipliers $\tilde{\lambda}_1, \dots, \tilde{\lambda}_{N_L+N_R}$.

In the first set of computations we study the importance of the scalar products for the numerical computations. In these computations the bounds on q were inactive. We use the weighted scalar products introduced in Section 6 and their corresponding norms. The scalar product $\langle \cdot, \cdot \rangle_{\mathcal{Q}_h}$ is used in the truncation criteria $\|C_q^* \tilde{\lambda}\|_{\mathcal{Q}_h} < 10^{-5}$ and, more importantly, for the computations of the BFGS updates. The

scalar product $\langle \cdot, \cdot \rangle_{\Lambda_h}$ is used to compute quantities like $\langle \tilde{\lambda}, C \rangle_{\Lambda_h}$. These computations are compared with the ones in which the discretized problem (6.7) is solved as a nonlinear programming problem in $\mathbb{R}^{N_L+N_R} \times \mathbb{R}^{N_L+N_R+1}$. The truncation criteria is $\|C_q^T \lambda\|_{\mathbb{R}^{N_L+N_R}} < 10^{-5}$.

First, we observe that the SQP method with weighted scalar products requires significantly fewer iterations to converge. The results are summarized in Table 7.1 in which we compare the two SQP versions for various choices of numbers of updates stored.

Table 7.1: Number of SQP iterations versus number L of updates stored ($N_L = 100, N_R = 100$).

Using weighted scalar products		Using Euclidean scalar products	
L	Iterations	L	Iterations
10	68	10	88
20	54	20	75
30	45	30	52
40	45	40	52

The superiority of the SQP method with weighted scalar products over the one with Euclidean scalar products, not only shows in the number of iterations, but also in the quality of the computed solution. Typical results are shown in Figures 7.1 and 7.2. The differences between computed velocity and target velocity and between computed area and the cubic area function are significantly larger for the computations using Euclidean scalar products. Moreover, the results computed using weighted scalar products and limited memory BFGS updates with $L = 30$ or $L = 40$ are virtually identical, whereas, the logarithmic derivatives of the area functions computed using Euclidean scalar products and $L = 30$ or $L = 40$ were significantly different. This shows that the relation between the infinite dimensional problem and its discretization is not only of theoretical interest, but also promises significant advantages from a computational point of view. As we have noted before, the reason for this behavior is that the underlying infinite dimensional problem dominates the discretized problems. If the discretized problems are treated as finite dimensional nonlinear programming problems, i.e., if the underlying infinite dimensional problem structure is ignored, then the problems often become artificially ill-conditioned and the performance of the optimization algorithms usually deteriorates as the discretization is refined. In our examples, an ill-conditioning is indicated by the oscillating parameter functions q_L, q_R shown in e.g. Figure 7.2. The use of weighted scalar products that are obtained from the discretization of the proper scalar products of the infinite dimensional problem take the underlying infinite dimensional problem structure into account. The resulting implementation of the SQP method is consistent with the formulation of the SQP method in the infinite dimensional framework.

The next results concern target data with errors. Although the target data \mathcal{d} was constructed using a different discretization for the area than the one used in the optimal design problem, the target data is almost feasible in the sense that we can find an area and a velocity profile such that $u \approx \mathcal{d}$. In the following we will use nonfeasible target data. This is done by modifying the procedure for the computation of the target data used previously. Now we compute two cubic splines using the points $(x_1, u_1), \dots, (x_{200-s}, u_{200-s})$ and $(x_{201+s}, u_{201+s}), \dots, (x_{400}, u_{400})$ and then we join these two cubic splines by constructing a cubic polynomial interpolating $(x_{200-s}, u_{200-s}), (x_{201+s}, u_{201+s})$ and the derivatives of the two previously constructed splines at x_{200+s} and x_{201+s} , respectively. This gives target data that are “smoother” around the shock.

Computations corresponding to the following figures were done using weighted scalar products and the

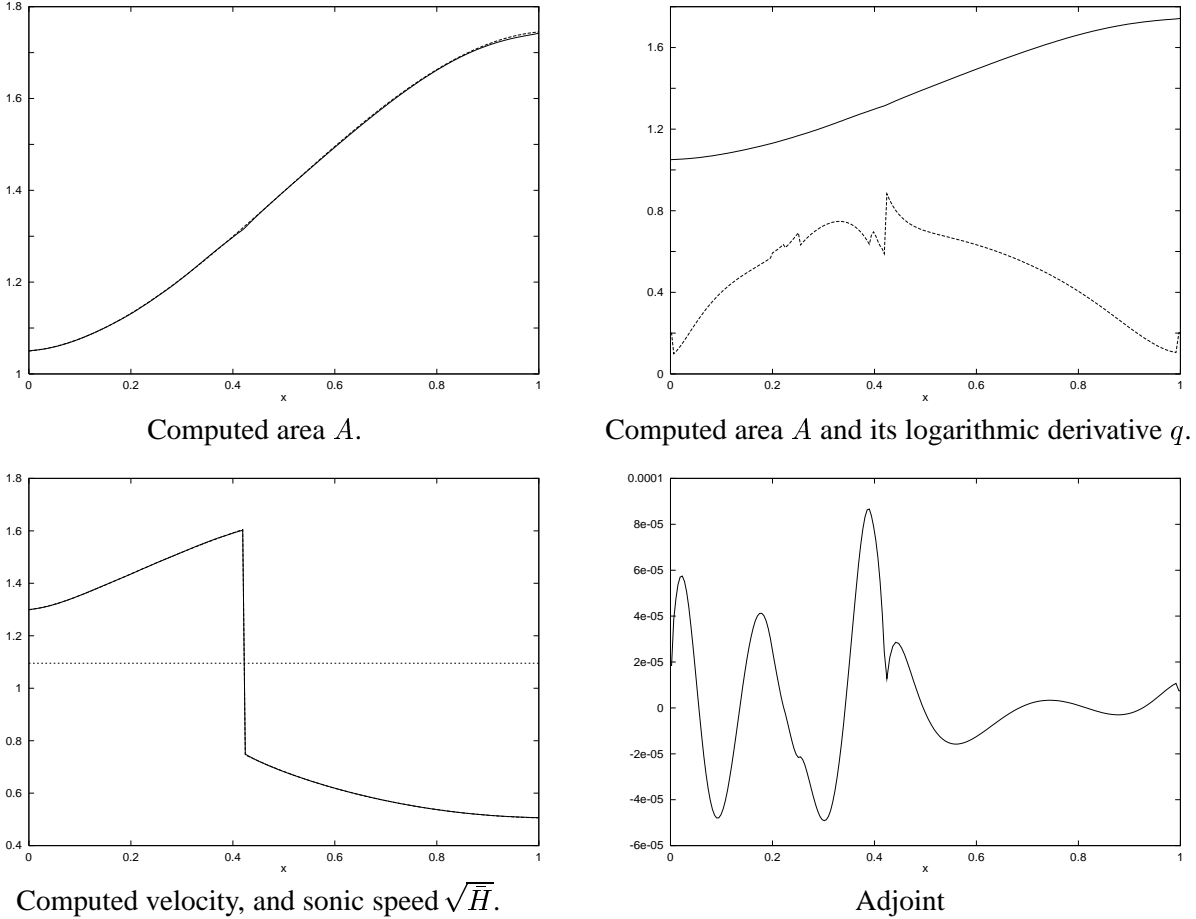


Figure 7.1: Computed area, velocity and adjoint using weighted scalar products and $L = 40$, $N_L = 100$, $N_R = 100$.

parameters $L = 40$ and $N_L = 100$, $N_R = 100$. The target data were computed using $s = 10$. The bounds $q_{\text{low}} = 0$ and $q_{\text{upp}} = 1$ were both active for some indices, as can be seen in Figure 7.3. The SQP method converged after 74 iterations. The function value at truncation was $J^h = 0.79 * 10^{-3}$, the norm of the constraints was $\|C^h\|_{\Lambda_h} = 0.75 * 10^{-7}$.

The fact that the logarithmic derivative is zero is due to the fact that the target velocity is not monotonically increasing left of the estimated shock location. In fact, if we consider the infinite dimensional problem, then the state equation (2.17) implies that

$$q_L = \frac{(1 - \bar{H}/u^2)u_x}{x_s(\bar{\gamma}u - \bar{H}/u)}.$$

If the target velocity u^d is decreasing left of the computed shock, then, in order to be close to u^d , the computed velocity tries to imitate the nature of the target flow and, hence, the logarithmic derivative q of the area tries to become negative. See the previous equation for q . Of course, the constraints prevent q_L from becoming negative. Similar reasoning can be used to explain why the logarithmic derivative of the area to the right of the computed shock, q_R is at its upper bound, q_{upp} .

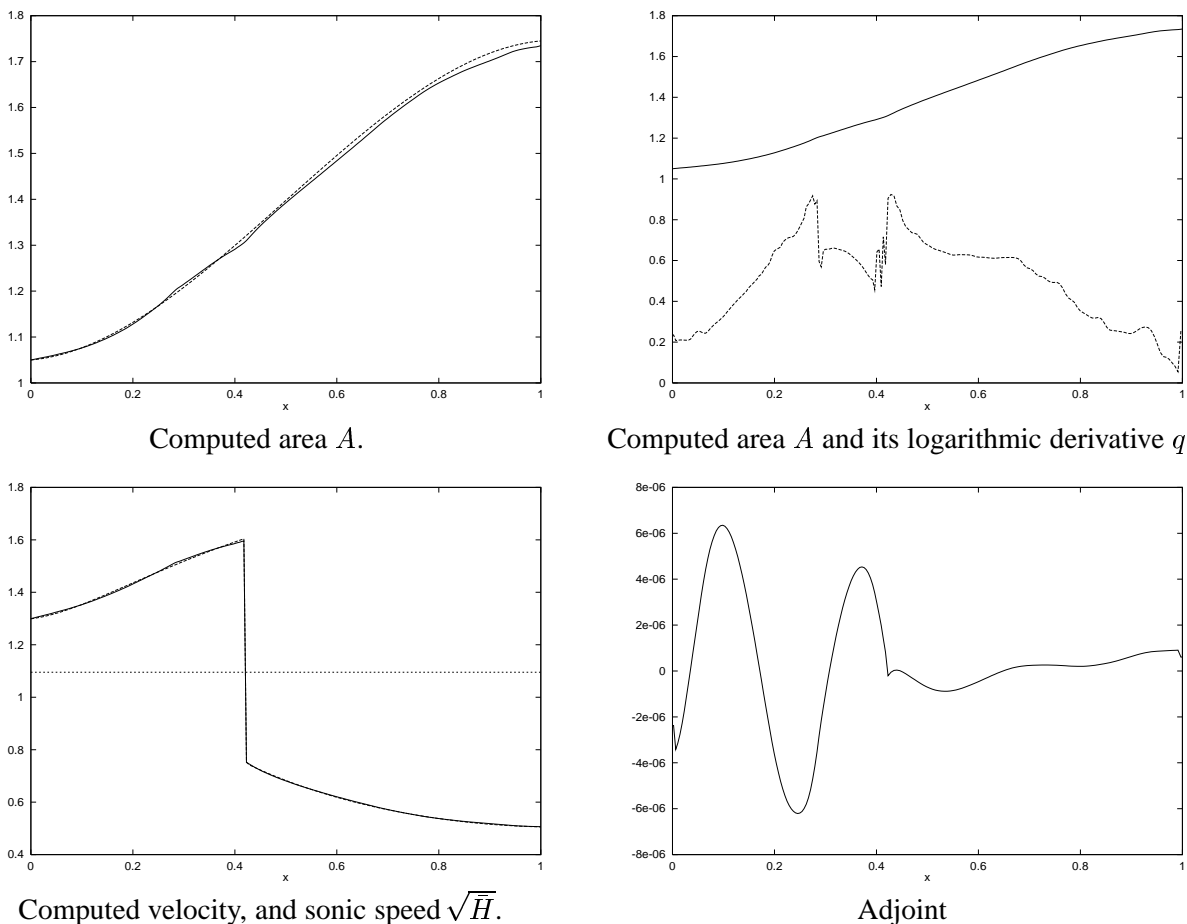


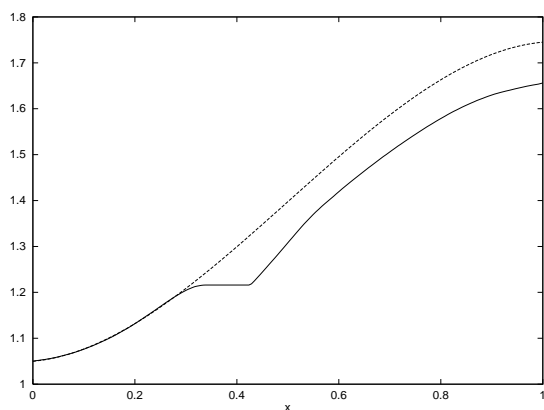
Figure 7.2: Computed area, velocity and adjoint using Euclidean scalar products and $L = 40$, $N_L = 100$, $N_R = 100$.

These results should not be surprising, as the lagrangian (6.20) is linear in the design variable, (q_L, q_R) , and optimal control theory tells us that this is a candidate for “bang-bang” control. In these cases, the bounds play an important role in the solution of the problem, as in most cases, a solution would not exist, without bounds. The region where the bound constraints on the design variables are inactive appears to correspond to a case of singular control, and we find the flows are perfectly matched in these regions. These results are similar to those obtained in [16]. In this case the Lagrange multipliers $\tilde{\lambda}_1, \dots, \tilde{\lambda}_{N_L+N_R}$ will generally not be zero in regions where the bounds are active, c.f. (5.11), (5.12). This behavior can be observed in Figure 7.3.

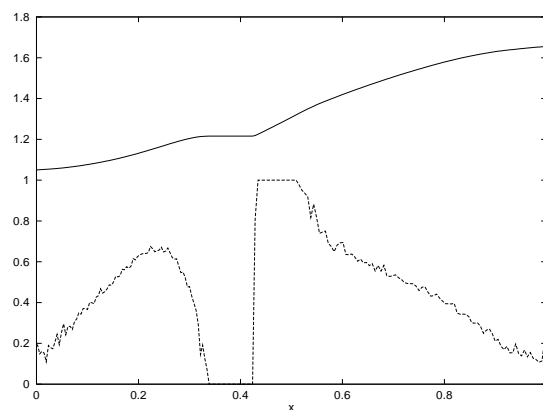
The presence of the lower bound at $q_{\text{low}} = 0$ is important in this case, for the results to make physical sense. The magnitude of the upper bound also seems to be important, as might be expected. We tried to run the same problem with $q_{\text{low}} = 0$ and $q_{\text{upp}} = 10$. However, the SQP algorithm stopped because the maximum number of iterations 100 was exceeded. The reason is that a spike evolves in the function q right of the estimated shock.

A similar situation prevails for the case where the discretization for the computed solution, N_L, N_R , exceeds the number of discrete grid points used to represent the target data. For the numerical example discussed above, when N_L or N_R exceeds 200, we find that some subintervals exist in the region around the

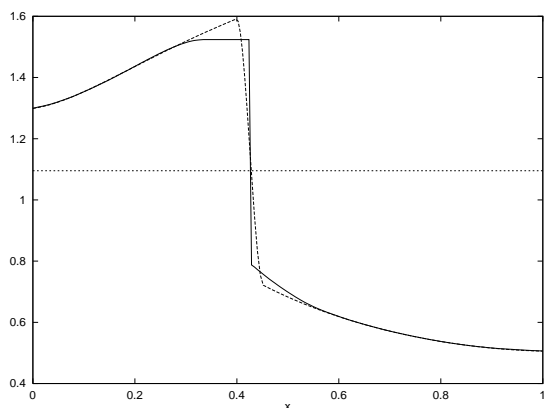
shock where the left and right target velocities are connected by a cubic spline. Since this cubic spline causes a smoothing, effects similar to those observed with the perturbed, smooth target data discussed previously were observed, for some cases. Thus, if a target velocity similar to the one used here has to be identified, then it seems to be important that the discretization of the problem is sufficiently coarse relative to the target data.



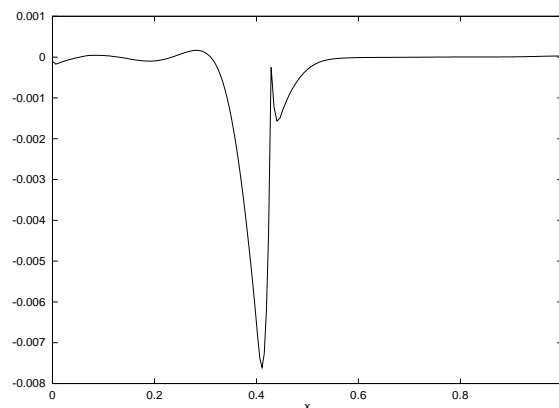
Computed versus 'exact' area A .



Computed area A and its logarithmic derivative q .



Computed velocity, and sonic speed \sqrt{H} .



Adjoint

Figure 7.3: Computed area, velocity, and adjoint using smoothed data.

8 Conclusions

In this paper we have studied a design optimization problem involving a compressible flow with a shock. The differentiability of the constraint functions and the formulation of the optimality conditions in the presence of shocks is a difficult issue. We used a formulation that treats the shock location as an explicit state variable. This allowed us to perform a rigorous mathematical analysis of the problem. We were able to sharply resolve the discontinuity while preserving differentiability of the map from design parameters to flow solution. Moreover, under suitable conditions we have established that the linearization of this map is invertible, that this inverse is uniformly bounded in a neighborhood of feasible points and that the usual

first order necessary optimality conditions are valid. An important finding of our study is that the co-state is discontinuous at the shock location, unless the target velocity can be matched perfectly.

The structure of the infinite dimensional problem is inherited by its discretization. However, important observations can be made concerning the numerical solution of the discretized optimal control problem. One can view the discretized optimal control problem as a nonlinear programming problem in $\mathbb{R}^{N_L+N_R} \times \mathbb{R}^{N_L+N_R+1}$. On the other hand, one can establish the relation between the original, infinite dimensional problem and its discretization. This leads to slight reformulations of the optimality conditions and the introduction of weighted constraints that correspond to the infinite dimensional formulation of the problem. This has been proven valuable for the numerical performance of the SQP algorithm used for the solution of the optimization problem. The use of weighted scalar products, i.e. the use of the infinite dimensional nature of the problem, reduced the number of iterations significantly and improved the quality of the computed solution.

Our results show that while a straightforward off-the-shelf application of SQP methods will likely fail, a careful analysis of the problem and an incorporation of the problem structure allows the successful application of these powerful methods. The extension of the result presented in this paper to the full Euler equations is part of our ongoing research.

References

- [1] W. Alt. The Lagrange–Newton method for infinite dimensional optimization problems. *Numer. Funct. Anal. Optim.*, 11:201–224, 1990.
- [2] W. Alt and K. Malanowski. The Lagrange–Newton method for nonlinear optimal control problems. *Computational Optimization and Applications*, 2:77–100, 1993.
- [3] J. D. Anderson. *Modern Compressible Flow with Historical Perspective*. McGraw–Hill Series in Aeronautical and Aerospace Engineering. McGraw–Hill, New York, St Louis, San Francisco, London, Paris, Tokyo, second edition, 1990.
- [4] J. T. Borggaard. *The Sensitivity Equation Method for Optimal Design*. PhD thesis, Virginia Polytechnic Institute and State University, Department of Mathematics, Blacksburg, VA 24061–0123, USA, 1994.
- [5] E. M. Cliff, M. Heinkenschloss, and A. Shenoy. An optimal design problem governed by the 1–d Euler equations. In *Proceedings from the 6th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, AIAA Paper 96–3993*, pages 114–121, Bellevue, Washington, 1996.
- [6] E. M. Cliff, M. Heinkenschloss, and A. Shenoy. An optimal control problem for flows with discontinuities. *Journal of Optimization Theory and Applications*, 94:273–309, 1997.
- [7] J. E. Dennis, M. Heinkenschloss, and L. N. Vicente. Trust–region interior–point algorithms for a class of nonlinear programming problems. *SIAM J. Control and Optimization*, 36:1750–1794, 1998.
- [8] P. D. Frank and G. R. Shubin. A comparison of optimization–based approaches for a model computational aerodynamics design problem. *J. Comput. Physics*, 98:74–89, 1992.
- [9] M. Heinkenschloss. Projected sequential quadratic programming methods. *SIAM J. Optimization*, 6:373–417, 1996.

- [10] M. Heinkenschloss and L. N. Vicente. Analysis of inexact trust–region interior–point SQP algorithms. Technical Report TR95–18, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005–1892, 1995. <http://www.caam.rice.edu/~heinken/papers/Papers.html>.
- [11] C. Hirsch. *Numerical Computation of Internal and External Flows, Volume 2: Computational Methods for Inviscid and Viscous Flows*. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1991.
- [12] D. M. Hwang and C. T. Kelley. Convergence of Broyden’s method in Banach spaces. *SIAM J. Optimization*, 2:505–532, 1992.
- [13] A. Iollo, M. D. Salas, and S. Ta’asan. Shape optimization governed by the Euler equations using an adjoint method. Technical Report 93–78, ICASE, NASA Langley Research Center, Hampton VA 23681–0001, 1993.
- [14] F.-S. Kupfer. An infinite dimensional convergence theory for reduced SQP methods in Hilbert space. *SIAM J. Optimization*, 6:126–163, 1996.
- [15] R. Narducci, B. Grossman, and R. T. Haftka. Sensitivity algorithms for an inverse design problem involving a shock wave. *Inverse Problems in Engineering*, 2:49–83, 1995.
- [16] A. Shenoy and E. Cliff. An optimal control formulation for a flow matching problem. In *Proceedings from the 5th AIAA/USAF/NASA/ISSMO Symposium On Multidisciplinary Analysis And Optimization, Panama City Beach, September 7-9*, pages 520–528, 1994.