SOLUTION OF OPTIMAL CONTROL PROBLEMS BY A POINTWISE PROJECTED NEWTON METHOD

C.T. KELLEY* AND E.W. SACHS†

Abstract. In the context of optimal control of ordinary differential equations, we prove local superlinear convergence and constraint identification results for an extension of the projected Newton method of Bertsekas. The estimates are also valid for discretized versions of the method-problem pair.

Key words. projected Newton iteration, optimal control

AMS(MOS) subject classifications. 47H17, 49K15, 49M15, 65J15, 65K10

1. Introduction. In many areas of optimal control the problems are formulated with simple constraints on the control. For this type of problems, the gradient projection type algorithms have proven to be quite successful, because they are able to take into account the structure of the underlying optimization problem. Another interesting feature of these methods is that they often can be formulated in infinite dimensional spaces which is important for the application to optimal control problems.

In general, let H denote a Hilbert space and for some closed convex subset $U \in H$ consider the optimization problem

$$\text{Minimize} \quad \phi(u) \quad \text{subject to} \quad u \in U. \tag{1.1}$$

If $\mathcal{P} : H \to U$ denotes the projection onto the feasible set, then the gradient projection method iterates are given by

$$u_+ = \mathcal{P}(u_c - \alpha_c \nabla \phi(u_c)) \tag{1.2}$$

where $\alpha > 0$ is determined by a step-size rule. In Hilbert space, this algorithm has been formulated and investigated by Goldstein [10] and Levitin and Polyak [14]. The books [4] and [3] discussed the convergence properties of gradient projection methods. In [7] a thorough convergence analysis of the gradient projection method was presented which yields various convergence rates of the algorithm under various assumptions.

Since the gradient projection method as presented in (1.2) is based on and identical for $U = H$ with the steepest descent method, its convergence properties exhibit locally a rather slow rate. This was the motivation to extend Newton's method to a projection method. There are basically two routes by which this goal can be achieved.

If one considers in the unconstrained case a Newton step as the solution of the minimization of a quadratic approximation of ϕ at the current iterates u_c, then for a problem of the type (1.1) one would have to solve

$$\text{Minimize} \quad (\nabla \phi(u_c), u - u_c) + \frac{1}{2}(u - u_c, \nabla^2 \phi(u_c)(u - u_c)) \quad \text{subject to} \quad u \in U. \tag{1.3}$$

This algorithm has been analyzed in [14] and [6]. The disadvantage of the method (1.3) is that at each step a quadratic problem with constraints needs to be solved. The simplicity of the constraints cannot be used in a direct way through the projection $\mathcal{P}$ except in solving the quadratic problem.

The other route to extend Newton's method to the constrained case is as follows: Instead of projecting the steepest descent direction onto the feasible set one projects the Newton direction onto the feasible set.

$$u_+ = \mathcal{P}(u_c - \alpha_c (\nabla^2 \phi(u_c))^{-1} \nabla \phi(u_c)) \tag{1.4}$$

* North Carolina State University, Department of Mathematics, Box 8205, Raleigh, N. C. 27695-8205. The research of this author was supported by National Science Foundation grant #DMS-9024622, Air Force Office of Scientific Research grant #AFOSR-FQ8671-9101094, and North Atlantic Treaty Organization grant #CRG 920067. Computing was partially supported by an allocation of time from the North Carolina Supercomputing Center.

† Universität Trier, FB IV – Mathematik, Postfach 3825, 5500 Trier, Federal Republic of Germany. The research of this author was supported by the Volkswagen-Stiftung and North Atlantic Treaty Organization grant #CRG 920067.

This method utilizes again the simple projection but has the drawback that it does not always produce a descent in the objective function. Bertsekas [1] and [2] introduced for the finite dimensional case with simple constraints such as upper and lower bounds on the variables a projected Newton method which alleviated this problem. For $H = R^n$ let

$$u_+ = \mathcal{P}(u_c - \alpha_c D_c^{-1} \nabla \phi(u_c)) \tag{1.5}$$

where D_c is a properly chosen matrix such that descent in the objective function is ensured. Let C_J denote the map which sets the components which lie in J of a vector $u \in R^n$ to zero. Then Bertsekas suggested that

$$D_c = C_J^T \nabla^2 \phi(u_c) C_J + C_{J^c}^T I C_{J^c}$$

where J contains the components of u_c which are active and the corresponding components of $\nabla \phi(u_c)$ point outside the feasible set. J^c denotes the complement of J in the index set. This algorithm combined with a proper step-size rule can be shown to converge locally at a quadratic rate since it identifies all active constraints after finitely many steps and becomes Newton's method for an unconstrained problem. The assumptions required for superlinear convergence of the method proposed in this paper include assumptions of the type of second order sufficiency 2.4 and of nondegeneracy 2.3. An approach similar to (1.3) might not need the latter assumption, however at the expense of a larger problem (1.3) to be solved at each step.

Since optimal control problems are problems formulated in function space, an analysis of the projected Newton method in this framework would give some insight for the case of fine discretizations. As shown in [12] this issue is important because the identification of finite indices is only mesh independent if proper measures are taken. The goal of this paper is to extend the algorithm to the infinite dimensional setting of optimal control problems.

The class of problems we seek to solve is

$$\text{minimize} \int_0^T L(x(t),\, u(t),\, t)\, dt$$

over $u \in U$ such that $x \in W_N^{1,\infty}[0,T]$ solves

$$\dot{x}(t) = f(x(t), u(t), t), \quad x(0) = x_0, \quad t \in (0,T).$$

Here $f : R^N \times R \times [0,T] \to R^N$ and $L : R^N \times R \times [0,T] \to R$ and for $t \in [0,T]$ we let U be given by

$$U(t) = \{u \in L^\infty[0,T] \,|\, u_{min}(t) \le u(t) \le u_{max}(t)\}$$

with u_{min} and u_{max} in $L^\infty[0,T]$.

The projection $\mathcal{P}$ is the map on $L^\infty[0,T]$ given by

$$\mathcal{P}(u)(t) = \begin{cases} u_{min}(t), & \text{if } u(t) \le u_{min}(t), \\ u(t), & \text{if } u_{min}(t) \le u(t) \le u_{max}(t), \\ u_{max}(t), & \text{if } u(t) \ge u_{max}(t), \end{cases} \quad . \tag{1.6}$$

We use the notation

$$L_N^\infty[0,T] = L^\infty([0,T]; R^N), \quad W_N^{1,\infty}[0,T] = W^{1,\infty}([0,T]; R^N)$$

for the spaces of R^N valued functions on $[0,T]$ having components in $L^\infty[0,T]$ and $W^{1,\infty}[0,T]$ respectively. If $w : [0,T] \to R^N$ has components w_i the norms on $L_N^\infty[0,T]$ and $W_N^{1,\infty}[0,T]$ are given by

$$\|w\|_{L_N^\infty[0,T]} = \sum_{i=1}^N \|w_i\|_{L^\infty[0,T]} \text{ and } \|w\|_{W_N^{1,\infty}[0,T]} = \sum_{i=1}^N \|w_i\|_{W^{1,\infty}[0,T]}.$$

The L^∞ and $W^{1,\infty}$ are defined by

$$\|u\|_{L^\infty[0,T]} = \text{ess-sup}_{t\in[0,T]}|u(t)| \text{ and } \|u\|_{W^{1,\infty}[0,T]} = \|u\|_{L^\infty[0,T]} + \|du/dt\|_{L^\infty[0,T]}$$

We will work in the spaces

$$X = W_N^{1,\infty}[0,T] \oplus W_N^{1,\infty}[0,T] \oplus L^\infty[0,T] \text{ and } Y = L_N^\infty[0,T] \oplus L_N^\infty[0,T] \oplus L^\infty[0,T].$$

For the unconstrained case, the first order necessary conditions for optimality can be defined in terms of the Hamiltonian function $H : R^N \times R^N \times R \times R \to R$

$$H(p,x,u,t) = f^T(x,u,t)\,p + L(x,u,t), \quad (p,x,u,t) \in R^{2N+2}$$

Usually, $p \in W_N^{1,\infty}[0,T]$ denotes the solution of the adjoint equation

$$-\dot{p} = f_x^T p + L_x^T \quad , \quad p(T) = 0.$$

For simplicity we will also use the notation for the Hamiltonian

$$H(p,x,u)(t) := H(p(t),x(t),u(t),t), \quad t \in [0,T]$$

and likewise for the partial derivatives of H.

The first order necessary conditions for the unconstrained case can be expressed as the system of nonlinear equations

$$F(z) = F(p,x,u) \;=\; \begin{pmatrix} \dot{x} - f(x,u,\cdot) \\ \dot{p} + H_x(p,x,u) \\ H_u(p,x,u) \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{1.7}$$

for $z = (p,x,u)^T \in X$ and $F : X \to Y$. The advantage of solving the system (1.7) over applying a Newton-like method directly to $\nabla f = H_u = 0$ is that the linear equations for the Newton steps in (1.7) can be expressed as linear equations for the new iterates without solving a nonlinear differential equation.

For the constrained case, the third equation in (1.7) must be modified to take the constraints into account. The system of nonlinear equations we consider here is

$$\mathcal{F}(z) = \mathcal{F}(p,x,u) = \begin{pmatrix} \dot{x} - f(x,u,\cdot) \\ \dot{p} + H_x(p,x,u) \\ u - \mathcal{P}(u - H_u(p,x,u)) \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{1.8}$$

To formulate the algorithm we introduce some more notation: For $I \subset [0,T]$, $A = [0,T] \setminus I$, and $z \in X$ define

$$Q_I z = (p, x, \chi_I u)^T.$$

$$\mathcal{F}_I(z) = Q_I F(p, x, \chi_I u + \chi_A u^*).$$

Note that in the definition of $\mathcal{F}_I$, $Q_I F$ and not $Q_I \mathcal{F}$ is used. This is important not only to make $\mathcal{F}_I'$ well defined but also for the success of the algorithm described in §4.

$\mathcal{F}_I$ can be regarded as a map from

$$X_I = W_N^{1,\infty}[0,T] \oplus W_N^{1,\infty}[0,T] \oplus L^\infty[I].$$

to

$$Y_I = L_N^\infty[0,T] \oplus L_N^\infty[0,T] \oplus L^\infty[I].$$

If the third component of $\mathcal{F}_I$ is understood to be identically zero on A we may also regard $\mathcal{F}_I$ as a map from X to Y. We will use $\mathcal{F}_I$ as both a theoretical and computational tool. When used in computation we must have knowledge of u^* on A. This is analogous to identification of the active set in finite dimensional problems [1], [2]. In the infinite dimensional setting considered in this paper, complete identification of A^* is not possible. However, as we will show in Section 4, a useful subset of A^* can be identified.

The authors extended in [13] Bertsekas' gradient projection method to constrained compact fixed point problems. It was combined with a multilevel algorithm of Atkinson and Brakhage and applied to a parabolic boundary control problem with simple bound constraints on the control. In this paper, we do not assume any compactness of the nonlinear map and analyze the resulting algorithm in infinite dimensions. As we will see in Section 4 relaxing the assumptions with regard to the compactness gives rise to an additional smoothing step in the algorithm so that a proper identification of the active set can be done at the subsequent iteration.

Section 3 contains a series of lemmas which are rather technical but lead to an important estimate in Theorem 3.5 in which the norm of the residual $\mathcal{F}$ can be used in an upper bound on the distance of the current iterate to the solution in the strong X-norm. The transfer from a projection based method which has its natural formulation in a Hilbert space like L^2 to a L^∞-type norm in X poses in the analysis of the convergence various difficulties. In another context this aspect has been the focus of other research activities, see e.g. [9], [8]. The estimate in Theorem 3.5 enables us to show a result on the identification of the set of active indices. It estimates the measure of the set on which the active set at the current iterate differs from the active set of the solution by the distance of the iterate from the solution in the X-norm. This result is the key for the convergence analysis of the algorithm.

In the following algorithm we set $u_c = u^*$ on $\bar{A}$, which is a well defined step by Lemma 3.7, and then apply a projected Newton iteration with $\bar{A}(z_c)$ as the active set. This yields an intermediate iterate $(x_{1/2}, p_{1/2})$ for the state and costate. Then a smoothing step for x and p is added which determines the set $\bar{A}(z_+)$ at the next iterate properly. The iteration is formally given by the following algorithm.

Algorithm `proj_newt`$(\mathcal{F}, z_c, z_+)$ *Choose* $\bar{p} \in (1/2, 1)$

1. *Compute*

$$\bar{A}(z_c) = \{t \,|\, |H_u(z_c)(t)| \geq \|\mathcal{F}(z_c)\|_Y^{\bar{p}}\}.$$

2. *Set* $u_c = u^*$ *on* $\bar{A}(z_c)$.
3. *Compute the projected Newton step*

$$s = -Q_{\bar{I}} \mathcal{F}'_{\bar{I}}(z_c)^{-1} Q_{\bar{I}} F(z_c).$$

4. *Set* $z_{1/2} = \bar{P}(z_c + s)$ *and* $u_+ = u_{1/2}$, *where*

$$\bar{P} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \mathcal{P} \end{pmatrix}$$

5. *Compute*

$$\begin{aligned} x_+(t) &= \int_0^t f(x_{1/2}(s), u_+(s), s)\ ds + x_0 \\ p_+(t) &= \int_t^T H_x(p_{1/2}(s), x_{1/2}(s), u_+(s))\ ds \end{aligned}$$

Apart from the smoothing step, this algorithm differs from the gradient projection method given by Bertsekas even in the finite dimensional case. To clarify this point consider a optimal control problem where in step (1.5) the new state x_+ in Bertsekas' algorithm has to be computed by solving the nonlinear system equation. In Algorithm `proj_newt` the correction for the new state is computed from a linearized version of the system equation, see e.g. (6.6).

In Section 4 we prove the local convergence of the iterates of the algorithm to the solution which is superlinear. The order of the rate of convergence is dependent on how stringently the identification of the

active set is carried out. The convergence is of q-order $2\bar{p}$ depending on the choice of $\bar{p} > 1/2$ in $\bar{A}(z)$ in the algorithm. The norm which we use in the convergence statement is the X-norm which is stronger than the Y-norm.

The contents of Section 5 relate the assumptions of the lemmas and theorems to other assumptions which have been used in the literature. In particular, the relation to second order sufficiency conditions is clarified. Also, the role of the growth conditions in [6] is indicated in our context.

Section 6 contains comments on the implementation of the algorithm. Furthermore, we present for an example numbers illustrating the convergence rate estimates of the convergence theorem. In particular, the dependence on $\bar{p}$ can be observed.

2. Notation. In this section we introduce some more notation, assumptions and immediate consequences of these.

Since U is given by

$$U(t) = \{u \in L^\infty[0,T] \,|\, u_{min}(t) \le u(t) \le u_{max}(t)\}$$

we define the two point set

$$\partial U(t) = \{u_{min}(t), u_{max}(t)\}.$$

We assume the existence of a solution to the first order necessary conditions.

ASSUMPTION 2.1. *There exists $z^* = (p^*, x^*, u^*)$ such that $\mathcal{F}(z^*) = 0$.*

Since $z^* \in X$, its range is contained in the bounded set

$$\mathcal{R} = \{\xi \in R^N \times R^N \times R \,|\, \|\xi\|_\infty \le \|z^*\|_X\}.$$

Also $(x,u) \in \mathcal{R}_1 \subset R^N \times R$ where

$$\mathcal{R}_1 = \{\zeta \in R^N \times R \,|\, \|\zeta\|_\infty \le \|z^*\|_X\}.$$

ASSUMPTION 2.2. *There is an open set $\mathcal{R}_0 \supset \mathcal{R}_1$ such that f, L and their first and second partial derivatives with respect to x and u are uniformly Lipschitz continuous on $\mathcal{R}_0 \times [0,T]$.*

An immediate consequence of Assumption 2.2 and the fact that for $z, w \in X$, $F'(z) - F'(w)$ is a multiplication operator and not a differential operator is the following lemma.

In the rest of the paper we denote function space norms by $\|\cdot\|$ and norms in R^k by

$$|x| = \max\{|x_j| \mid j = 1, ..., k\}.$$

LEMMA 2.1. *There are $\sigma^*, L_F, M_F > 0$ such that for all $z, w \in \mathcal{N} = \{v \in X \,|\, \|v - z^*\|_X < \sigma^*\}$ and $t \in [0,T]$,*

$$|F(z)(t) - F(w)(t)| \le M_F |z(t) - w(t)|, \tag{2.1}$$

$$\begin{aligned} &\|F'(z) - F'(w)\|_{\mathcal{L}(X,Y)} \le L_F \|z - w\|_Y \text{ and} \\ &\|F'(z) - F'(w)\|_{\mathcal{L}(Y,Y)} \le L_F \|z - w\|_Y. \end{aligned} \tag{2.2}$$

We define active and inactive sets for u by

$$A(u) = \{t \,|\, u(t) \in \partial U(t)\} \text{ and } I(u) = [0,T] \setminus A(u) = \{t \,|\, u(t) \in \operatorname{int}(U)(t)\}. \tag{2.3}$$

In (2.3) $\operatorname{int}(U)(t)$ is defined to be the interval

$$\operatorname{int}(U)(t) = (u_{min}(t), u_{max}(t)).$$

We let

$$A^* = A(u^*) \text{ and } I^* = I(u^*).$$

If $S \subset R^k$ for some k and $t \in R^k$ we denote the distance from t to S by

$$\text{dist}(t, S) = \inf\{s \in S \,|\, |t - s|\}.$$

As in [13] we make structural assumptions on the active set at the solution.

ASSUMPTION 2.3. *There is $\nu \in (0,1)$ such that $u_{max}(t) \geq u_{min}(t) + \nu$ for all $t \in [0,T]$. A^* is the closure of a finite union of open sets. On each component of A^* either $u = u_{max}$ or $u = u_{min}$.*

Moreover, there is c_1 such that

$$\begin{aligned} |H_u(p^*, x^*, u^*)(t)| &\geq c_1 \, dist(t, \partial A^*) \text{ for all } t \in A^* \text{ and} \\ dist(t, A^*) &\leq c_1^{-1} \, dist(u^*(t), \partial U(t)) \text{ for all } t \in I^*. \end{aligned} \tag{2.4}$$

This assumption is a condition on the slope of the function H_u which is also called switching function in the control context. In Section 5 we relate Assumption 2.3 to the growth condition in [7]. We also give a connection to a similar condition in [12] which has been used for the finite identification of active indices in the gradient projection method.

The fact that we consider problems of dimension one with regard to the control together with the assumption on the structure of A^* yields the following lemma, which we give without proof.

LEMMA 2.2. *There is $c_0 > 0$ such that for all $\delta > 0$ the sets*

$$E_\delta = \{t \in R \,|\, dist(t, \partial A^*) < \delta\}$$

are uniformly bounded in measure by

$$\mu(E_\delta) \leq c_0 \delta. \tag{2.5}$$

We define projections P_1 and P_2 for $z = (p, x, u)^T \in Y$ by

$$P_1 z = (p, x, 0)^T, \; P_2 z = (0, 0, u)^T.$$

The observations that $P_1 F = P_1 \mathcal{F}$ and $Q_J P_1 = P_1 Q_J = P_1$ for any $J \subset [0,T]$ lead to the following result.

PROPOSITION 2.3. *Let Assumptions 2.1 and 2.2 hold. If $I \subset I^*$ then $\mathcal{F}_I(z^*) = 0$ and for all $t \in [0,T]$*

$$|P_1 \mathcal{F}_I(z)(t)| \leq \|P_1 \mathcal{F}(z)\|_Y + M_F |(I - Q_I)(z - z^*)(t)|.$$

Proof. Since

$$P_1 \mathcal{F}_I(z) = P_1 Q_I F(Q_I z + (I - Q_I) z^*)$$

and

$$P_1 Q_I \mathcal{F}(z) = P_1 Q_I F(z)$$

we have the result by Lemma 2.1. □

Nonsingularity assumptions are also complicated by the constraints.

ASSUMPTION 2.4. *There are K_η and $\tilde{\eta} > 0$ such that if*

$$I \subset \{t \,|\, |H_u(p^*, x^*, u^*)(t)| \leq \tilde{\eta}\} \tag{2.6}$$

then $\mathcal{F}_I'(z^*)$ is a nonsingular map from X_I to Y_I. Moreover,

$$\|\mathcal{F}_I'(z^*)^{-1}\|_{\mathcal{L}(Y_I,X_I)} \le K_\eta \tag{2.7}$$

and for any measurable set $S \subset [0,T]$

$$\|P_1\mathcal{F}_I'(z^*)^{-1}(I-Q_S)\|_{\mathcal{L}(Y_I,Y_I)} \le K_\eta \mu(S). \tag{2.8}$$

There is $\tilde{\alpha}$ such that

$$H_{uu}(p^*,x^*,u^*)(t) \ge \tilde{\alpha}$$

for all $t \in I$.

This assumption is related to second order sufficiency conditions for optimal control problems. Details are discussed in Section 5.

3. Identification of the Active Set. We begin by summarizing the definitions of the many projections that will be used in the following sections. Beginning with the definition of $\mathcal{P}$ in (1.6) as the projection of u onto the feasible set we defined $\bar{P}\ z = (p,x,u)^T \in Y$ by

$$\bar{P}(z) = (p, x, \mathcal{P}(u))^T.$$

For a measurable set $I \in [0,T]$ we defined

$$Q_I(z) = (p, x, \chi_I u)^T$$

where χ_I is the characteristic function of the set I. Finally we define P_1 and P_2, which decompose z into the state-costate and control parts by

$$P_1(z) = (p,x,0)^T,\ P_2 z = (0,0,u)^T.$$

For $z \in X$ we let

$$e = (e_p, e_x, e_u)^T = z - z^*.$$

We want to show that provided $\|e\|_X$ is sufficiently small, it can be estimated by a constant multiple of $\|\mathcal{F}\|_Y$. This estimate is important in that it allows us to identify a subset of A^*. We will require a sequence of lemmas.

LEMMA 3.1. *Assume that Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Then there are B_0, c_2, and $\tau_0 > 0$ such that if $z = \bar{P}(z) \in X$ is such that $\|e\|_X \le \tau_0$ then for $S(z) = \{t \mid u(t) - H_u(p,x,u)(t) \notin U(t)\} \cap I^* \cap I(u)$*

$$|P_2\mathcal{F}_{I(u)\cap I^*}(z)(t)| \le B_0(\|\mathcal{F}(z)\|_Y + \chi_{S(z)}(t)\|e\|_Y),$$

where

$$\mu(S(z)) \le c_2\|e\|_Y.$$

Proof. We assume that $\tau_0 < \min(\sigma^*, \nu)$, where σ^* is the diameter of the set $\mathcal{N}$ in Lemma 2.1. We set $J = I^* \cap I(u)$ in this proof and set $\|e\|_Y = \sigma \le \|e\|_X$.

Observe that

$$P_2\mathcal{F}_J(z) = (0,0,\chi_J H_u(p,x,\chi_J u + \chi_{J^c}u^*))^T = (0,0,\chi_J H_u(p,x,u))^T$$

and therefore

$$P_2\mathcal{F}_J(z) = (0,0,(1-\chi_{S(z)})\chi_J H_u(p,x,u) + \chi_{S(z)}H_u(p,x,u)(t))^T. \tag{3.1}$$

On $S(z)$ we have $H_u(p^*, x^*, u^*) = 0$, since $S(z) \subset J \subset I^*$. Hence for $t \in S(z)$

$$|\chi_{S(z)} H_u(p, x, u)(t)| = |\chi_{S(z)}(H_u(p, x, u) - H_u(p^*, x^*, u^*))(t)| \leq \chi_{S(z)} M_F \|e\|_Y. \tag{3.2}$$

For $t \in S(z)^c$ we have $u - H_u(p, x, u) \in U$ and

$$(1 - \chi_{S(z)})(0, 0, H_u(p, x, u))^T = (1 - \chi_{S(z)}) P_2 \mathcal{F}(z)$$

and hence

$$|(1 - \chi_{S(z)}) H_u(p, x, u)| \leq \|\mathcal{F}\|_Y.$$

This proves the first part of the assertion with $B_0 = \max\{1, M_F\}$.

We now complete the proof with an estimate of the measure of $S(z)$. Let $t \in S(z)$. The estimate (3.2) and $H_u(p^*, x^*, u^*)(t) = 0$ yield

$$\|\chi_{S(z)}(u^* - H_u(p^*, x^*, u^*) - (u - H_u(p, x, u)))\|_\infty \leq (1 + M_F)\sigma$$

Since $u - H_u(p, x, u) \notin U(t)$ we must have

$$\operatorname{dist}(u^*(t), \partial U(t)) \leq (1 + M_F)\sigma$$

Hence, by Assumption 2.3,

$$\operatorname{dist}(t, A^*) \leq c_1^{-1}(1 + M_F)\sigma$$

for any $t \in S(z)$. Therefore

$$S(z) \subset \{t \,|\, \operatorname{dist}(t, \partial A^*) \leq c_1^{-1}(1 + M_F)\sigma\},$$

and hence, by Lemma 2.2,

$$\mu(S(z)) \leq c_0 c_1^{-1}(1 + M_F)\sigma.$$

This completes the proof with $c_2 = c_0 c_1^{-1}(1 + M_F)$. □

COROLLARY 3.2. *Assume that Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Then if* $\|e\|_X \leq \tau_0$

$$\|P_1 \mathcal{F}'_{I(u)\cap I^*}(z^*)^{-1} P_2 \mathcal{F}_{I(u)\cap I^*}(z)\|_Y \leq B_1(\|\mathcal{F}(z)\|_Y + \|e\|_Y^2). \tag{3.3}$$

Proof. The result follows directly from Assumption 2.4 and Lemma 3.1 with $B_1 = K_\eta B_0(1 + c_2)$. □

LEMMA 3.3. *Assume that Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Then there are* B_5 *and* $\tau_1 > 0$ *such that if* $z = \bar{P}(z) \in X$ *is such that* $\|e\|_X \leq \tau_1$ *then*

$$\|P_1 \mathcal{F}'_{I(u)\cap I^*}(z^*)^{-1} P_1 \mathcal{F}_{I(u)\cap I^*}(z)\|_Y \leq B_5(\|\mathcal{F}(z)\|_Y + \|e\|_Y^2). \tag{3.4}$$

Proof. We assume that $\tau_1 \leq \tau_0$ with τ_0 chosen in Lemma 3.1. We set $I = I(u)$ and $A = A(u)$ in this proof. We let $\|e\|_Y = \sigma \leq \|e\|_X \leq \tau_1$ and $\|\mathcal{F}(z)\|_Y = \delta$. We let $J = I^* \cap I$.

By Proposition 2.3

$$|P_1 \mathcal{F}_J(z)| \leq \delta + M_F(1 - \chi_J)|e_u| = \delta + M_F \chi_{S_U}|e_u|,$$

for all $t \in [0, T]$, where S_U is the support of $(1 - \chi_J)e_u$. Our next task is to study S_U.

Assumption 2.3 implies that if we choose $\tau_1 < \nu$ then $u = u^*$ on $E = A(u) \cap A^*$. Hence $(1 - \chi_J)e_u$ is nonzero only in the set

$$E_1 = ([0,T] \setminus (E \cup J)) = (I^* \cap A) \cup (I \cap A^*)$$

and hence $S_U \subset E_1$. We consider three cases. First, if $t \in (I^* \cap A)$ then $H_u(p^*, x^*, u^*) = 0$. If $t \in (I \cap A^*)$ and $u - H_u(p, x, u) \in \text{int}(U)$, then $|H_u(p, x, u)(t)| = |P_2\mathcal{F}(z)(t)| \leq \delta$ and therefore

$$|H_u(p^*, x^*, u^*)(t)| \leq \delta + M_F\sigma.$$

If we now let

$$E_2 = \{t \in E_1 \,|\, u - H_u(p, x, u) \in \text{int}(U)\} \cup (I^* \cap A)$$

then for all $t \in E_2$ and $B_2 = 1 + M_F$

$$|H_u(p^*, x^*, u^*)(t)| \leq B_2(\delta + \sigma). \tag{3.5}$$

We must consider a third case, that for

$$t \in E_3 = \{t \in I \cap A^* \,|\, u - H_u(p, x, u) \notin \text{int}(U)\}.$$

Since

$$|\mathcal{P}(u - H_u(p, x, u)) - u^*| = |\mathcal{P}(u - H_u(p, x, u)) - \mathcal{P}(u^* - H_u(p^*, x^*, u^*))| \leq (1 + M_F)\sigma \tag{3.6}$$

for all $t \in [0, T]$, we may reduce τ_1 if needed so that $(1 + M_F)\tau_1 < \nu$ to conclude that if $\sigma < \tau_1$ and if $t \in E_3$ then

$$u^* = \mathcal{P}(u - H_u(p, x, u))$$

and hence

$$P_2\mathcal{F}(z) = (0, 0, u - \mathcal{P}(u - H_u(p, x, u)))^T = (0, 0, e_u)^T. \tag{3.7}$$

This implies that

$$|\chi_{E_3} e_u| \leq \|\mathcal{F}(z)\|_Y = \delta.$$

At this point we have

$$|P_1\mathcal{F}_J(z)| \leq \delta + M_F(1 - \chi_J)|e_u| = \delta + M_F\chi_{E_2}|e_u| + M_F\chi_{E_3}|e_u| \leq (1 + M_F)\delta + M_F\chi_{E_2}|e_u|. \tag{3.8}$$

The Banach Lemma, Assumption 2.4, and Lemma 2.1 imply that if $\sigma < \sigma^*$ then

$$\|{\mathcal{F}'_J}^{-1}\|_{\mathcal{L}(Y_I, X_I)} \leq K_\eta/(1 - \sigma M_F).$$

Hence, reducing τ_1 if needed so that $\tau_1 M_F < 1/2$, we have

$$\|(\mathcal{F}'_J)^{-1} P_1\mathcal{F}_J(z)\|_Y \leq 2K_\eta((1 + M_F)\delta + M_F\sigma\mu(E_2)). \tag{3.9}$$

We now estimate the measure of E_2. Using (2.4) we see that if $t \in E_2$ then either $t \in I \cap A^*$ and from (3.5)

$$\text{dist}(t, \partial A^*) \leq c_1^{-1} B_2(\sigma + \delta)$$

or $t \in I^* \cap A$ and

$$\operatorname{dist}(t, \partial A^*) \le c_1^{-1} \operatorname{dist}(u^*(t), \partial U(t)) \le c_1^{-1} |u(t) - u^*(t)| \le c_1^{-1} \sigma.$$

Hence, setting $B_3 = c_1^{-1} \max(B_2, 1)$

$$\operatorname{dist}(t, \partial A^*) \le c_1^{-1} B_3 (\sigma + \delta) \text{ for all } t \in E_2.$$

Then Lemma 2.2 implies that

$$\mu(E_2) \le B_4(\sigma + \delta) \tag{3.10}$$

where $B_4 = B_3 c_0 c_1^{-1}$. If we reduce τ_1 if needed so that $M_F B_4 \tau_1 \le 1$ and set $B_5 = 2K_\eta(2 + M_F + M_F B_4)$ then the proof is complete. □

LEMMA 3.4. *Assume that Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Then there are B_9 and $\tau_2 > 0$ such that if $z = \bar{P}(z) \in X$ is such that $\|e\|_X \le \tau_2$ then*

$$\|P_2 e\|_Y \le B_9(\|\mathcal{F}(z)\|_Y + \|P_2 e\|_Y^2 + \|P_1 e\|_Y). \tag{3.11}$$

Proof. We assume that $\tau_2 \le \tau_1$ with τ_1 from Lemma 3.3. Let $e_p = p - p^*$, $e_x = x - x^*$, and $e_u = u - u^*$. We set $I = I(u)$ and $A = A(u)$ in this proof. We let $\|e\|_Y = \sigma \le \|e\|_X \le \tau_1$, $\|P_1 e\|_Y = \beta$, and $\|\mathcal{F}(z)\|_Y = \delta$. We let $J = I^* \cap I$.

We define a set $\mathcal{S}$ by

$$\mathcal{S} = \{t \mid u - H_u(p, x, u) \in U(t)\}.$$

Note that if $t \in \mathcal{S}$ then the third component of $\mathcal{F}(z)$ is H_u. Therefore $|H_u(p, x, u)| \le \delta$ for all $t \in \mathcal{S}$. Therefore, since $H_u(p^*, x^*, u^*) = 0$ on I^*, both $H_u(p^*, x^*, u^*) = 0$ and $H_u(p, x, u) = O(\delta)$ on the set $\mathcal{S} \cap I^*$. Hence, for all $t \in \mathcal{S} \cap I^*$ we have

$$\begin{aligned} 0 = H_u(p^*, x^*, u^*) &= H_u(p, x, u^*) + O(\beta) \\ &= H_u(p, x, u) - H_{uu}(p^*, x^*, u^*) e_u + O(\sigma^2 + \beta). \end{aligned}$$

The bound away from 0 of H_{uu} implies that there is B_6 such that

$$\|\chi_{\mathcal{S} \cap I^*} e_u\|_\infty \le B_6(\delta + \sigma^2 + \beta).$$

On $\mathcal{S} \cap A^*$, $|H_u(p, x, u)| \le \delta$ and hence

$$H_u(p^*, x^*, u^*) + H_{uu}(p^*, x^*, u^*) e_u = O(\delta + \sigma^2 + \beta).$$

As for $H_u(p^*, x^*, u^*)$ we know that

$$H_u(p^*, x^*, u^*)(u - u^*) \ge 0.$$

If $u^* = u_{max}$, say, then $e_u \le 0$ and therefore $H_u(p^*, x^*, u^*) \le 0$. Hence

$$0 \le -H_u(p^*, x^*, u^*) = H_{uu}(p^*, x^*, u^*) e_u + O(\delta + \sigma^2 + \beta).$$

Since $e_u \le 0$ we must have $|e_u| = O(\delta + \sigma^2 + \beta)$. Applying a similar argument to the case where $u = u_{min}$ implies that

$$\|\chi_{\mathcal{S} \cap A^*} e_u\|_\infty \le B_7(\delta + \sigma^2 + \beta).$$

We must now estimate $|e_u|$ on $\mathcal{S}^c = [0,T] \setminus \mathcal{S}$. If $t \in \mathcal{S}^c$ we have $u - H_u(p,x,u) \notin U(t)$ and therefore $\mathcal{P}(u - H_u(p,x,u)) \in \partial U(t)$. On $A^* \cap \mathcal{S}^c$, $u^* = \mathcal{P}(u - H_u(p,x,u))$ if σ_0 is sufficiently small by (3.6). Therefore (3.7) holds and so

$$\|\chi_{A^* \cap \mathcal{S}^c} e_u\|_\infty \leq \delta.$$

On $I^* \cap \mathcal{S}^c$, $0 = H_u(p^*, x^*, u^*)$. Since $\mathcal{P}(u - H_u(p,x,u)) \in \partial U(t)$, either $\mathcal{P}(u - H_u(p,x,u)) = u_{max}$ or $\mathcal{P}(u - H_u(p,x,u)) = u_{min}$. If $\mathcal{P}(u - H_u(p,x,u)) = u_{max}$, say, then

$$|u - \mathcal{P}(u - H_u(p,x,u))| = |u - u_{max}| \leq \delta. \tag{3.12}$$

Hence

$$u_{max} + \delta - H_u(p,x,u) \geq u - H_u(p,x,u) \geq u_{max}$$

and therefore $H_u(p,x,u) \leq \delta$. Similarly if $\mathcal{P}(u - H_u(p,x,u)) = u_{min}$, $H_u(p,x,u) \geq -\delta$. Now, if $\mathcal{P}(u - H_u(p,x,u)) = u_{max}$ then $H_u(p,x,u) \leq \delta$ and therefore

$$-\delta \leq u - \mathcal{P}(u - H_u(p,x,u)) = u - u_{max} \leq u - u^*. \tag{3.13}$$

Hence $e_u \geq -\delta$. Also,

$$\begin{aligned} 0 = H_u(p^*, x^*, u^*) &= H_u(p,x,u) - H_{uu}(p^*, x^*, u^*) e_u + O(\delta + \sigma^2 + \beta) \\ &\leq -H_{uu}(p^*, x^*, u^*) e_u + \delta + O(\delta + \sigma^2 + \beta) \\ &= -H_{uu}(p^*, x^*, u^*) e_u + O(\delta + \sigma^2 + \beta) \end{aligned} \tag{3.14}$$

Since $H^*_{uu} \geq \tilde{\alpha}$ on I^* by Assumption 2.4, (3.14) implies

$$e_u \leq \tilde{\alpha}^{-1} O(\delta + \sigma^2 + \beta) = O(\delta + \sigma^2 + \beta). \tag{3.15}$$

We may use (3.13) and (3.15) to conclude that

$$-\delta \leq e_u \leq O(\delta + \sigma^2 + \beta)$$

and therefore $e_u = O(\delta + \sigma^2 + \beta)$. The estimate is exactly the same if $u = u_{min}$. Hence there is B_8 such that

$$\|\chi_{I^* \cap \mathcal{S}^c} e_u\|_\infty \leq B_8(\delta + \sigma^2 + \beta).$$

This completes the proof with $B_9 = \max\{B_6, B_7, B_8\}$. □

THEOREM 3.5. *Assume that Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Then there are $K_X, \sigma_0 > 0$ such that if $z = \bar{P}(z) \in X$ satisfies $\|z - z^*\|_X \leq \sigma_0$ then*

$$\|z - z^*\|_X \leq K_X \|\mathcal{F}(z)\|_Y. \tag{3.16}$$

Proof. We assume that $\sigma_0 < \tau_1$ with τ_1 by Lemma 3.3. We let $\|e\|_Y = \sigma \leq \|e\|_X \leq \sigma_0$ and $\|\mathcal{F}(z)\|_Y = \delta$. We let $J = I^* \cap I(u)$. By Proposition 2.3 $\mathcal{F}_J(z^*) = 0$.

$$\begin{aligned} \mathcal{F}_J(z) &= \mathcal{F}_J(z^*) + \int_0^1 \mathcal{F}'_J(z^* + tQ_J e) Q_J e \, dt \\ &= \mathcal{F}'_J(z^*) Q_J e + \int_0^1 (\mathcal{F}'_J(z^* + tQ_J e) - \mathcal{F}'_J(z^*)) Q_J e \, dt. \end{aligned} \tag{3.17}$$

We now have, using (3.17),

$$Q_J e = \mathcal{F}_J'(z^*)^{-1}\left(\mathcal{F}_J(z) - \int_0^1 (\mathcal{F}_J'(z^* + tQ_J e) - \mathcal{F}_J'(z^*))Q_J e\, dt\right). \tag{3.18}$$

By Lemma 2.1 the integral term satisfies

$$\left\| \int_0^1 (\mathcal{F}_J'(z^* + tQ_J e) - \mathcal{F}_J'(z^*))Q_J e\, dt \right\|_Y \leq L_F \sigma^2/2 \tag{3.19}$$

where L_F is the bound in Lemma 2.1 and hence with (2.7)

$$\left\| \mathcal{F}_J'(z^*)^{-1} \int_0^1 (\mathcal{F}_J'(z^* + tQ_J e) - \mathcal{F}_J'(z^*))Q_J e\, dt \right\|_X \leq K_\eta L_F \sigma^2. \tag{3.20}$$

It remains to estimate $\mathcal{F}_J'(z^*)^{-1}\mathcal{F}_J(z)$.

By Lemma 3.3 and Corollary 3.2 we have

$$\|P_1 \mathcal{F}_J'(z^*)^{-1}\mathcal{F}_J(z)\|_Y \leq B_{10}(\delta + \sigma^2), \tag{3.21}$$

where $B_{10} = B_1 + B_5$. At this point we can estimate (3.18) using (3.20), (3.21) and $\|x\|_Y \leq \|x\|_X$ for $x \in X$ as follows

$$\|P_1 e\|_Y = \|P_1 Q_J e\|_Y \leq B_{11}(\delta + \sigma^2), \tag{3.22}$$

where $B_{11} = B_{10} + K_\eta L_F$.

By the definition of P_2, Lemma 3.4 and Assumption 2.4 we have

$$\|P_2 e\|_X = \|P_2 e\|_Y \leq B_9(1 + B_{11})(\delta + \sigma^2). \tag{3.23}$$

From this we conclude with (3.22) that

$$\sigma = \|e\|_Y \leq B_{12}(\delta + \sigma^2)$$

where $B_{12} = B_{11} + B_9(1 + B_{11})$. Hence, reducing σ_0 if necessary so that $B_{12}\sigma \leq B_{12}\sigma_0 \leq 1/2$,

$$\|e\|_Y = \sigma \leq 2B_{12}\delta \text{ and } \sigma^2 \leq \sigma 2B_{12}\delta \leq \delta. \tag{3.24}$$

In order to obtain an estimate for $\|e\|_X$, not $\|e\|_Y$, we use (3.23) and the second two parts of (3.24)

$$\|e\|_X \leq \|P_1 e\|_X + \|P_2 e\|_X \leq \|P_1 Q_J e\|_X + 2B_9(1 + B_{11})\delta. \tag{3.25}$$

We estimate $\|P_1 Q_J e\|_X$ in two parts based on (3.18). Lemma 3.1 and Proposition 2.3 together with (3.24) imply that

$$\begin{aligned} \|\mathcal{F}_J(z)\|_Y &\leq \|P_1\mathcal{F}(z)\|_Y + M_F\|(I - Q_J)e\|_Y + B_0(\|\mathcal{F}(z)\|_Y + \|e\|_Y) \\ &\leq (1 + B_0)\delta + (M_F + B_0)\|e\|_Y \leq B_{13}\delta \end{aligned}$$

where

$$B_{13} = (1 + 2M_F B_{12}) + B_0(1 + 2B_{12}).$$

Therefore by Assumption 2.4

$$\|\mathcal{F}_J'(z^*)^{-1}\mathcal{F}_J(z)\|_X \leq K_\eta \|\mathcal{F}_J(z)\|_Y \leq K_\eta B_{13}\delta. \tag{3.26}$$

Hence we can conclude the proof by estimating (3.25) further with (3.18), (3.26), (3.20) and (3.24) to obtain (3.16) with $K_X = K_\eta(B_{13} + L_F) + 2B_9(1 + B_{11})$. ∎

As a consequence we have a result on identification of the active set A^*.

THEOREM 3.6. *Assume that Assumptions 2.1, 2.2, 2.3, and 2.4 hold. For all $\bar{p} \in (0,1)$ there is σ_1 such that if $\|z - z^*\|_X < \sigma_1, z \neq z^*$ and*

$$\bar{A}(z) = \{t \,|\, |H_u(z)(t)| \geq \|\mathcal{F}(z)\|_Y^{\bar{p}}\} \tag{3.27}$$

then $\bar{A}(z) \subset A^*$ and there is c_μ such that

$$\mu(A^* \setminus \bar{A}(z)) \leq c_\mu \|z - z^*\|_X^{\bar{p}}. \tag{3.28}$$

Proof. Let $\sigma_1 < \sigma_0$ as in Theorem 3.5, so that the consequences of Theorem 3.5 hold. Let $\|\mathcal{F}(z)\|_Y = \delta$ and $\|z - z^*\|_X = \sigma$. Let L_H denote the Lipschitz constant of H_u. For $t \in \bar{A}(z)$ we have

$$|H_u(z^*)(t)| \geq |H_u(z)(t)| - L_H\sigma \geq \delta^{\bar{p}} - L_H\sigma \geq (K_X^{-1}\sigma)^{\bar{p}} - L_H\sigma > 0$$

if

$$\sigma^{1-\bar{p}} \leq K_X^{-\bar{p}}/L_H.$$

Hence if

$$\sigma_1 \leq (K_X^{-\bar{p}}/L_H)^{1/(1-\bar{p})}$$

then $t \in \bar{A}(z)$ implies that $H_u(z^*)(t) > 0$ and therefore that $t \in A^*$. We now set

$$\sigma_1 = \min(\sigma_0, (K_X^{-\bar{p}}/L_H)^{1/(1-\bar{p})}).$$

If $t \in A^* \setminus \bar{A}$ then, using (2.4) from Assumption 2.3,

$$\begin{aligned} c_1 \mathrm{dist}(t, \partial A^*) \leq |H_u(z^*)(t)| \quad &\leq |H_u(z^*)(t) - H_u(z)| + |H_u(z)| < L_H\sigma + \delta^{\bar{p}} \\ &\leq L_H\sigma + M_F^{\bar{p}}\sigma^{\bar{p}}. \end{aligned} \tag{3.29}$$

This implies that $t \in E_\zeta$ where

$$\zeta = (L_H\sigma + M_F^{\bar{p}}\sigma^{\bar{p}})/c_1.$$

Hence

$$\mu(A^* \setminus \bar{A}(z)) \leq \mu(E_\zeta) \leq c_0\zeta.$$

If we set

$$c_\mu = c_0(L_H\sigma_1^{1-\bar{p}} + M_F^{\bar{p}})/c_1$$

the proof is complete. ∎

The final result in this section is that if z is sufficiently near z^* then u can be set to u^* on $\bar{A}(z)$ in a well defined way.

LEMMA 3.7. *Assume that Assumptions 2.1, 2.2, 2.4, and 2.3 hold. Then for all $\bar{p} \in (0,1)$ there is σ_2 such that if $\|z - z^*\|_X < \sigma_2, z \neq z^*$ and $t \in \bar{A}(z)$ then*

$$|u(t) - u^*(t)| < |u(t) - w(t)|$$

for all $w \neq u^*$, $w(t) \in \partial U(t)$. Therefore the assignment of u to the nearest of u_{min} or u_{max} on $\bar{A}$ is well defined and decreases the X norm of $z - z^*$.

Proof. Let $\sigma_2 \leq \sigma_1$ so that the conclusions of Theorem 3.6 hold. If $t \in \bar{A}(z) \subset A^*$ then either $u^*(t) = u_{max}(t)$ or $u^*(t) = u_{min}(t)$. Without loss of generality we assume that $u^*(t) = u_{max}(t)$. Letting $\|z - z^*\|_X = \sigma$ we have $|u(t) - u_{max}(t)| \leq \sigma$ and

$$|u(t) - u_{min}(t)| \geq |u^*(t) - u_{min}(t)| - |u(t) - u^*(t)| \geq \nu - \sigma,$$

where ν is from Assumption 2.3. This completes the proof if $\sigma_2 < \nu/2$. ∎

4. The Algorithm. Let z be such that the conclusions of Theorem 3.5 hold and let $\|\mathcal{F}(z)\|_Y = \delta$. Let $\bar{p} \in (0,1)$ and let $\bar{A}(z)$ be given by (3.27). Let $\bar{I} = \bar{A}^c = [0,T] \setminus \bar{A}$.

Note that if $v = Q_{\bar{I}} z + (I - Q_{\bar{I}})(z - \mathcal{F}(z))$ then $v_u = u^*$ on $\bar{A}(z)$ and hence $\mathcal{F}_I$ can be computed since the value of z^* on $\bar{A}$ is known.

The variant of the projected Newton algorithm that we propose here makes the transition from a current iterate z_c to a new point z_+ by setting $u_c = u^*$ on $\bar{A}$, which is a well defined step by Lemma 3.7, and then applying a projected Newton iteration with $\bar{A}(z_c)$ as the active set. The iteration is formally given by the following algorithm.

ALGORITHM 4.1. *Algorithm* `proj_newt`$(\mathcal{F}, z_c, z_+)$

1. Compute

$$\bar{A}(z_c) = \{t \,|\, |H_u(z_c)(t)| \geq \|\mathcal{F}(z_c)\|_Y^{\bar{p}}\}.$$

2. Set $u_c = u^*$ on $\bar{A}(z_c)$.

3. Compute the projected Newton step

$$\begin{aligned} s &= -(I - Q_{\bar{I}})\mathcal{F}(z_c) - Q_{\bar{I}}\mathcal{F}'_{\bar{I}}(z_c)^{-1} Q_{\bar{I}} \mathcal{F}_{\bar{I}}(z_c) \\ &= -Q_{\bar{I}}\mathcal{F}'_{\bar{I}}(z_c)^{-1} Q_{\bar{I}} F(z_c) \qquad (4.1) \\ &= -(Q_{\bar{I}} F'(z_c) Q_{\bar{I}})^{-1} Q_{\bar{I}} F(z_c). \end{aligned}$$

4. Set $z_{1/2} = \bar{P}(z_c + s)$ and $u_+ = u_{1/2}$, where

$$\bar{P} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \mathcal{P} \end{pmatrix}$$

5. Compute

$$\begin{aligned} x_+(t) &= \int_0^t f(x_{1/2}(s), u_+(s), s)\ ds + x_0 \\ p_+(t) &= \int_t^T H_x(p_{1/2}(s), x_{1/2}(s), u_+(s))\ ds \end{aligned}$$

Remark: The term $(I - Q_{\bar{I}})\mathcal{F}(z_c)$ in the right side of (4.1) in step 3 of Algorithm 4.1 vanishes because of the change in u_c on $\bar{A}$ in step 2 in the algorithm. We include it in the first line of (4.1) to emphasize the similarity of the algorithm we propose with the projected Newton algorithm in [2]. Hence after the overwriting of u_c with u^* on $\bar{A}$ in step 2 we need only compute the step on $\bar{I}$. Another effect of step 2 is the relation

$$Q_{\bar{I}} \mathcal{F}_{\bar{I}}(z_c) = Q_{\bar{I}} F(z_c),$$

which follows from $u = u^*$ on $\bar{A}$.

The convergence behavior of the iteration is given by the following.

THEOREM 4.1. *Let the assumptions of Theorem 3.5 hold. There are $K_N > 0$ and $\sigma_3 > 0$ such that if $\|e_c\|_X < \sigma_3$ and z_+ is given by Algorithm 4.1 then*

$$\|e_+\|_X \leq K_N \|e_c\|_X^{2\bar{p}}. \qquad (4.2)$$

Proof. First let $\sigma_3 \leq \sigma_2$ so that the conclusions of our previous results are valid. The proof begins by estimating $\|P_1 e_+\|_X$ in terms of $\|e_c\|_X$ and $\|P_2 e_+\|$. This first step reduces the proof to an estimate of $\|P_2 e_+\|_X$ in terms of $\|e_c\|$, which only involves the u-component of the error.

Step 5 of Algorithm 4.1 serves as a smoothing step. By definition we have $\|e_+\|_X = \|e_+\|_Y + \|\frac{d}{dt}P_1 e_+\|_Y$. Assumption 2.2 yields with a constant $B_{14} > 0$

$$\|\frac{d}{dt}P_1 e_+\|_Y = \|f(x_{1/2}, u_+) - f(x^*, u^*)\|_{L^\infty} + \|H_x(p_{1/2}, x_{1/2}, u_+) - H_x(p^*, x^*, u^*)\|_{L^\infty} \le B_{14}\|e_{1/2}\|_Y$$

Similarly,

$$\begin{aligned} e_{x+}(t) &= \int_0^t (f(x_{1/2}(s), u_+(s)) - f(x^*(s), u^*(s)))ds \quad \text{and} \\ e_{p+}(t) &= \int_t^T H_x(p_{1/2}(s), x_{1/2}(s), u_+(s)) - H_x(p^*(s), x^*(s), u^*(s)))ds. \end{aligned} \tag{4.3}$$

Hence there is B_{15} such that

$$\|P_1 e_+\|_Y \le B_{15}\|e_{1/2}\|_Y.$$

Therefore

$$\|e_+\|_X \le (B_{14} + B_{15})\|e_{1/2}\|_Y.$$

Step 2 of Algorithm 4.1 forces $u_c = u^*$ on $\bar{A} \subset A^*$ by Theorem 3.6. $Q_{\bar{I}}s$ can be viewed as a perturbation of the Newton step $\bar{s}$ for the map $Q_{\bar{I}}(\mathcal{F}_{\bar{I}}(z) - \mathcal{F}_{\bar{I}}(z^*))$ from the point z_c. We have

$$\bar{s} = -Q_{\bar{I}}\mathcal{F}'_{\bar{I}}(z_c)^{-1}Q_{\bar{I}}(\mathcal{F}_{\bar{I}}(z_c) - \mathcal{F}_{\bar{I}}(z^*))$$

and $\bar{z}_{1/2} = z_c + \bar{s}$ satisfies

$$\|Q_{\bar{I}}\bar{e}_{1/2}\|_X \le B_{16}\|Q_{\bar{I}}e_c\|_X^2. \tag{4.4}$$

Here, for this proof,

$$B_{16} = K_\eta L_F/2$$

as is standard for Newton's method.

Now,

$$Q_{\bar{I}}e_{1/2} = Q_{\bar{I}}\bar{e}_{1/2} + \mathcal{F}'_{\bar{I}}(z_c)^{-1}\mathcal{F}_{\bar{I}}(z^*). \tag{4.5}$$

To estimate $\mathcal{F}'_{\bar{I}}(z_c)^{-1}\mathcal{F}_{\bar{I}}(z^*)$ we note that because $\bar{A} \subset A^*$, $I^* \subset \bar{I}$, $\mathcal{F}_{\bar{I}}(z^*)$ vanishes on I^*. $\mathcal{F}_{\bar{I}}(z^*)$ also vanishes on $\bar{A} \subset A^*$ by definition. Therefore

$$\mathcal{F}_{\bar{I}}(z^*) = P_2\mathcal{F}_{\bar{I}}(z^*) = \begin{pmatrix} 0 \\ 0 \\ \chi_{\bar{I}\setminus I^*}H_u(p^*, x^*, u^*) \end{pmatrix}.$$

Hence, as in (3.9), we may require σ_3 to be small enough so that

$$\|P_1\mathcal{F}'_{\bar{I}}(z_c)^{-1}\mathcal{F}_{\bar{I}}(z^*)\|_Y \le 2K_\eta\|\mathcal{F}_{\bar{I}}(z^*)\|_Y\,\mu(\bar{I}\setminus I^*). \tag{4.6}$$

We may estimate $\|\mathcal{F}_{\bar{I}}(z^*)\|_Y$ by using the fact that

$$\mathcal{F}_{\bar{I}}(z^*) = P_2\mathcal{F}_{\bar{I}}(z^*)$$

and the definition of $\bar{A}$ to find

$$\begin{aligned} \|\mathcal{F}_{\bar{I}}(z^*)\|_Y &\le \|\mathcal{F}_{\bar{I}}(z^*) - \mathcal{F}_{\bar{I}}(z_c)\|_Y\|\mathcal{F}_{\bar{I}}(z_c)\|_Y \\ &\le M_F\|e_c\|_X + M_F^{\bar{p}}\|e_c\|_Y^{\bar{p}} \\ &\le (M_F\sigma_2^{1-\bar{p}} + M_F^{\bar{p}})\|e_c\|_X^{\bar{p}}. \end{aligned}$$

The fact that $\bar{I} \setminus I^* = A^* \setminus \bar{A}$ and Theorem 3.6 imply

$$\mu(\bar{I} \setminus I^*) \le c_\mu \|e_c\|_X^{\bar{p}}. \tag{4.7}$$

From (4.5), (4.6), (4.7) we obtain

$$\|P_1 e_{1/2}\|_Y = \|P_1 Q_{\bar{I}} e_{1/2}\|_Y \le \|P_1 Q_{\bar{I}} \bar{e}_{1/2}\|_Y + \|P_1 \mathcal{F}'_{\bar{I}}(z_c)^{-1} \mathcal{F}_{\bar{I}}(z^*)\|_Y \tag{4.8}$$

(4.9)

$$\le B_{16} \|e_c\|_X^2 + K_\eta c_\mu (M_F \sigma_2^{1-\bar{p}} + M_F^{\bar{p}}) \|e_c\|_X^{2\bar{p}} \le B_{17} \|e_c\|_X^{2\bar{p}} \tag{4.10}$$

where

$$B_{17} = B_{16} \sigma_2^{2-2\bar{p}} + c_\mu K_\eta (M_F \sigma_2^{1-\bar{p}} + M_F^{\bar{p}}).$$

This gives

$$\|P_1 e_+\|_X \le B_{15} B_{17} \|e_c\|_X^{2\bar{p}} + \|P_2 e_+\|_X. \tag{4.11}$$

We must now consider the equation for the third component $P_2 z_+$ of the projected Newton iterate. If $t \in \bar{A} \subset A^*$ then $u_+ = u_c = u^*$ and the third components of the step, the current error, and the new error vanish.

By (4.1), on $\bar{I}$ we have

$$f_u(x_c, p_c, u_c) s_p + H_{ux}(x_c, p_c, u_c) s_x + H_{uu}(p_c, x_c, u_c) s_u = -H_u(p_c, x_c, u_c).$$

By Taylor's theorem and the fact that $f_u = H_{up}$ we have

$$H_{uu}(p_c, x_c, u_c) s_u = -(H_u(p_c, x_c, u_c) + H_{up}(p_c, x_c, u_c) s_p + H_{ux}(p_c, x_c, u_c) s_x) = -H_u(p_{1/2}, x_{1/2}, u_c) - \Delta_1,$$

where

$$\begin{aligned} \Delta_1 \quad &= \textstyle\int_0^1 (H_{up}(p_c + t s_p, x_c + t s_x, u_c) - H_{up}(p_c, x_c, u_c)) s_p \\ &\quad + (H_{ux}(p_c + t s_p, x_c + t s_x, u_c) - H_{ux}(p_c, x_c, u_c)) s_x \, dt. \end{aligned}$$

By Assumption 2.2

$$\|\Delta_1\|_\infty \le 2 M_H \|P_1 s\|_X^2$$

where M_H is an upper bound for the Lipschitz constants of $f_u = H_{up}$, H_{ux}, and H_{uu}.

Now,

$$H_{uu}(p_c, x_c, u_c) s_u = -H_u(p^*, x^*, u_c) + \Delta_2,$$

where

$$\Delta_2 = H_u(p^*, x^*, u) - H_u(p_{1/2}, x_{1/2}, u) - \Delta_1.$$

Since

$$\|P_1 s\|_X \le K_\eta \delta \le K_\eta L_F \|e_c\|_X$$

and (4.8) implies

$$\|H_u(p_{1/2}, x_{1/2}, u) - H_u(p^*, x^*, u)\|_\infty \le L_H \|P_1 e_{1/2}\|_Y \le L_H B_{17} \|e_c\|_X^{2\bar{p}},$$

we have Δ_2 can be bounded by

$$\|\Delta_2\|_\infty \leq B_{18}\|e_c\|_X^{2\bar{p}},$$

where

$$B_{18} = 2M_H K_\eta^2 L_F^2 \sigma_0^{2-2\bar{p}} + L_H B_{17}.$$

We expand $H_u(p^*, x^*, u_c)$ about u^* and apply Taylor's theorem again to obtain

$$H_u(p^*, x^*, u) = H_u(p^*, x^*, u^*) + H_{uu}(p^*, x^*, u^*)e_u + O(\|e_c\|_Y^2)$$

and hence

$$H_{uu}(p_c, x_c, u_c)s_u = -H_u(p^*, x^*, u^*) - H_{uu}(p^*, x^*, u^*)e_u + \Delta_3, \tag{4.12}$$

where

$$\|\Delta_3\|_\infty \leq B_{19}\|e_c\|_X^{2\bar{p}},$$

for some $B_{19} > 0$.

Let $\tilde{u}_+ = u_c + s_u$. Equation (4.12) may be rewritten as

$$\tilde{u}_+ = u^* - (H_{uu}(p_c, x_c, u_c))^{-1}H_u(p^*, x^*, u^*) - [1 - (H_{uu}(p_c, x_c, u_c))^{-1}H_{uu}(p^*, x^*, u^*)]e_u + \Delta_3. \tag{4.13}$$

Since

$$(1 - H_{uu}(p_c, x_c, u_c))^{-1}H_{uu}(p^*, x^*, u^*))e_u = O(\|e_c\|_Y^2)$$

we have, for all $t \in \bar{I}$,

$$\tilde{u}_+ = u^* - (H_{uu}(p_c, x_c, u_c))^{-1}H_u(p^*, x^*, u^*) + \Delta_4, \tag{4.14}$$

where

$$\|\Delta_4\|_\infty \leq B_{20}\|e_c\|_X^{2\bar{p}},$$

for some $B_{20} > 0$.

Since $H_{uu}(p_c, x_c, u_c) > \tilde{\alpha}$ for all $t \in \bar{I}$ we have that

$$u^* = \mathcal{P}(u^* - (H_{uu}(p_c, x_c, u_c))^{-1}H_u(p^*, x^*, u^*)),$$

for all $t \in \bar{I}$. Therefore

$$\begin{aligned} u_+ = \mathcal{P}(\tilde{u}_+) \quad &= \mathcal{P}(u^* - (H_{uu}(p_c, x_c, u_c))^{-1}H_u(p^*, x^*, u^*) + \Delta_4) \\ &= u^* + \Delta_5, \end{aligned}$$

where

$$\Delta_5 = \mathcal{P}(u^* - (H_{uu}(p_c, x_c, u_c))^{-1}H_u(p^*, x^*, u^*) + \Delta_4) - u^*$$

satisfies

$$\|\Delta_5\|_\infty \leq \|\Delta_4\|_\infty \leq B_{20}\|e_c\|_X^{2\bar{p}}.$$

Hence

$$\|P_2 e_+\|_X \le B_{20}\|e_c\|_X^{2\bar{p}}. \tag{4.15}$$

We combine this with (4.11) to obtain

$$\begin{aligned}\|e_+\|_X &\le \|P_1 e_+\|_X + \|P_2 e_+\|_X \\ &\le B_{15}B_{17}\|e_c\|_X^{2\bar{p}} + 2\|P_2 e_+\|_X \\ &\le (B_{15}B_{17} + 2B_{20})\|e_c\|_X^{2\bar{p}}.\end{aligned}$$

Setting $K_N = B_{15}B_{17} + 2B_{20}$ completes the proof. □

5. Assumptions. In this section we review the assumptions posed in Section 2 and relate them to other conditions used in the context of optimal control problems with ordinary differential equations.

Since the theory developed uses a L^∞-framework in contrast to L^2, there is no problem in establishing the proper differentiability assumptions of the mappings. Recall that $\mathcal{F}_I : X_I \to Y_I$ for some measurable set $I \subset [0,T]$ with $A = [0,T] \setminus I$ is defined by

$$\mathcal{F}_I(z) = Q_I F \begin{pmatrix} p \\ x \\ \chi_I u + \chi_A u^* \end{pmatrix} = \begin{pmatrix} \dot{x} - f(x, \chi_I u + \chi_A u^*, t) \\ \dot{p} + H_x(x, \chi_I u + \chi_A u^*, t) \\ \chi_I(H_u(x, \chi_I u + \chi_A u^*, t)) \end{pmatrix} \tag{5.1}$$

for $z \in X_I$. Therefore the derivative is given by

$$\mathcal{F}_I'(z)(\zeta) = Q_I F' \begin{pmatrix} p \\ x \\ \chi_I u + \chi_A u^* \end{pmatrix} \begin{pmatrix} \pi \\ \xi \\ \chi_I \nu \end{pmatrix} = \begin{pmatrix} \dot{\xi} - f_x \xi - f_u \chi_I \nu \\ \dot{\pi} + f_x^T \pi + H_{xx}\xi + H_{xu}\chi_I \nu \\ \chi_I(f_u^T \pi + H_{ux}\xi + H_{uu}\chi_I \nu) \end{pmatrix} \tag{5.2}$$

where $\zeta = (\pi, \xi, \nu) \in X$. In (5.2) we have omitted the arguments for the derivatives of the functions.

The regularity assumptions in Assumption 2.4 are related to second order sufficiency conditions in optimal control. In the papers [16] and [15] second order sufficiency conditions of the following type are used. A strengthened Legendre-Clebsch condition is posed with the existence of a solution to a Riccati equation, both appropriately altered to the case of control constraints. We assume the existence of a solution $Z(t) \in R^{n\times n}$ on [0,T] of the Riccati equation

$$-\dot{Z} = Z f_x + f_x^T Z + H_{xx} - (H_{xu} + Z f_u) H_{uu}^+ (H_{ux} + f_u^T Z), \quad Z(T) = 0 \tag{5.3}$$

where H_{uu}^+ is defined as

$$H_{uu}^+ = \chi_I H_{uu}^{-1} \chi_I.$$

LEMMA 5.1. *Let $z^* = (p^*, x^*, u^*) \in X$ be given. Assume that for I given by (2.6) there exists a solution $Z \in W_{n\times n}^{1,\infty}[0,T]$ of (5.3) and for some $\delta > 0$*

$$H_{uu}(t) \ge \delta \quad \text{a. e. on } I. \tag{5.4}$$

Then (2.7) and (2.8) of Assumption 2.4 hold.

Proof. For given $(a,b,c)^T \in Y^\infty$ let γ be the solution of the initial value problem

$$\dot{\gamma} - (-f_x^T + H_{xu}H_{uu}^+ f_u^T)\gamma = b - H_{xu}H_{uu}^+ c - Z(a + f_u H_{uu}^+ c), \quad \gamma(T) = 0. \tag{5.5}$$

With γ known, denote by ξ the solution of

$$\dot{\xi} + (f_u H_{uu}^+ f_u^T Z - f_x + f_u H_{uu}^+ f_u^T)\xi = a + f_u H_{uu}^+ c - f_u H_{uu}^+ f_u^T \gamma, \quad \xi(0) = 0. \tag{5.6}$$

Define π by

$$\pi = Z\xi + \gamma \tag{5.7}$$

and ν on I by

$$\nu = H_{uu}^{+}(c - f_u^T\pi - H_{ux}\xi), \tag{5.8}$$

which can be defined also as a function on $[0,T]$ by extension with 0.

Then one can verify that $(\pi,\xi,\nu)^T \in X_I$ solve the system

$$\mathcal{F}_I'(z^*)(\pi,\xi,\nu)^T = (a,b,c)^T, \tag{5.9}$$

which proves the surjectivity of $\mathcal{F}_I'(z^*)$. The continuous dependence of solutions of initial value problems on the right hand side of the differential equation allows us to deduce from (5.5) with (5.4) that $\|\gamma\|_{W^{1,\infty}}$ depends continuously on $\|(a,b,c)\|_{Y_I}$. Using this fact one obtains the same statement from (5.6) for ξ. Finally, one estimates the L^∞–norm of ν by (5.8) and this altogether yields the estimate

$$\|\mathcal{F}_I'(z^*)^{-1}(a,b,c)\|_{X_I} = \|(\pi,\xi,\nu)\|_{X_I} \le B_{21}\|(a,b,c)\|_{Y_I} \tag{5.10}$$

for some positive number B_{21}. This proves (2.7).

In order to show (2.8), let S be a measurable subset of $[0,T]$ and let $(a,b,c)^T \in Y^\infty$. Then

$$(I - Q_{S^c})(a,b,c)^T = (0,0,\chi_S c)$$

and let $(\pi,\xi,\nu)^T \in X_I$ be the solution of

$$\mathcal{F}_I'(z^*)(\pi,\xi,\nu)^T = (0,0,\chi_S c)^T = (I - Q_{S^c})(a,b,c)^T. \tag{5.11}$$

We define γ and ξ as solutions of (5.5) and (5.6), resp., with (a,b,c) replaced by $(0,0,\chi_S c)$. Then we obtain from the modified (5.5) that for some constant B_{22} we have

$$\|\gamma\|_\infty \le B_{22}\|\chi_S\nu\|_1 \le B_{22}T\mu(S)\|\nu\|_\infty$$

and a similar estimate follows from modified (5.6) for $\|\xi\|_\infty$. Hence we obtain with (5.11)

$$\|P_1\mathcal{F}_I'(z^*)^{-1}(I - Q_{S^c})(a,b,c)\|_{Y_I} = \|P_1(\pi,\xi,\nu)\|_{Y_I} = \|(\pi,\xi,0)\|_{Y_I} \le B_{23}\mu(S)\|\nu\|_\infty$$

which implies (2.8). □

The last condition in Assumption 2.4 is a trivial consequence of the assumption on the strengthened Legendre-Clebsch condition (5.4) in Lemma 5.1 if we choose I properly. We can relax the second order sufficiency conditions to hold only on I^* under a proper smoothness assumption on the control.

LEMMA 5.2. *Let $z^* = (p^*,x^*,u^*) \in X$ be given such that $u^* \in C[0,T]$ and let Assumption 2.3 hold. Assume that for I^* we have for some $\delta > 0$*

$$H_{uu}(t) \ge \delta \quad \text{a. e. on } I^* \tag{5.12}$$

there and that there exists a solution $Z \in W^{1,\infty}_{n\times n}[0,T]$ of (5.3) with $H_{uu}^{+} = \chi_{I^*}H_{uu}^{-1}\chi_{I^*}$. Then (2.7) and (2.8) of Assumption 2.4 hold.

Proof. Note that p^*, x^* are continuous as solutions of differential equations. With the assumption on u^* we have $H_{uu} \in C[0,T]$. Furthermore, Assumption 2.3 yields, that for small ρ

$$I_\rho = \{t \in [0,T] \;\mid\; |H_u(p^*,x^*,u^*)| \le \rho\}$$

we have that

$$\lim_{\rho\to 0}\mu(I_\rho \setminus I^*) = 0.$$

Since by (5.12) the continuous function H_{uu} is greater or equal $\delta > 0$ on I^*, we can choose $\rho = \tilde{\sigma} > 0$ small enough so that for an appropriately small $\tilde{\alpha}$

$$H_{uu}(t) \geq \tilde{\alpha} > 0 \text{ on } I_{\tilde{\sigma}}.$$

We have assumed the existence of a solution of the Riccati equation where H_{uu}^+ has support only on I^*. If $\tilde{\sigma}$ is small enough then the Riccati equation with H_{uu}^+ and support on $I_{\tilde{\sigma}}$ also has a bounded solution on $[0,T]$. This completes the proof. □

Next we discuss the statements in Assumption 2.3. For the finite dimensional case, a typical nondegeneracy condition would require that each component of $H_u(p^*,x^*,u^*)_i$ is nonzero if the corresponding component u_i^* of the optimal control lies in the active set A^*. The additional difficulty occuring for the infinite dimensional problem is that H_u can approach zero in different ways. Here we had to impose a requirement that $|H_u|$ grows at a similar rate as the distance from the boundary of the active set when moving away from the boundary. Obviously, this condition reduces to the previously mentioned nondegeneracy condition in finite dimensions.

We can relate (2.4) to a condition on the zeroes of the switching function H_u. A similar condition was used in [5, Theorem 6.2] and also in [17, (2.14)]. To state this more precisely we prove the following lemma.

LEMMA 5.3. *Assume that for* $0 \leq t_1 \leq \ldots \leq t_{2r+1} \leq T$

$$\{t \in [0,T] \mid H_u(x^*,p^*,u^*)(t) = 0\} = \bigcup_{i=0}^{r} [t_{2i}, t_{2i+1}] \tag{5.13}$$

and that the function $g(t) := |H_u(x^*,p^*,u^*)(t)|$ *is continuous and has one sided derivatives with*

$$g'_-(t_{2i}), g'_+(t_{2i+1}) > 0 \quad \text{for } i = 1,..,r. \tag{5.14}$$

Then the first line of (2.4) *in Assumption 2.3 holds.*

Proof. The assumption (5.13) yields that A^* consists of finitely many subintervals. Let $A_i^* = [t_{2i-1}, t_{2i}]$ denote such an interval. Then (5.14) implies that there are $\epsilon, m > 0$ such that

$$g(t) \geq m|t - t_{2i-1}| \text{ on } [t_{2i-1}, t_{2i-1} + \epsilon] \text{ and } g(t) \geq m|t - t_{2i}| \text{ on } [t_{2i} - \epsilon, t_{2i}].$$

By definition g is positive on (t_{2i-1}, t_{2i}). Therefore, we can choose $\epsilon > 0$ so small that

$$g(t) \geq m\epsilon \text{ on } [t_{2i-1} + \epsilon, t_{2i} - \epsilon].$$

With $m^* = \min\{m, 2\epsilon/(t_{2i} - t_{2i-1})\}$ we obtain

$$g(t) \geq m|t - t_{2i-1}| \text{ on } [t_{2i-1}, (t_{2i-1} + t_{2i})/2] \text{ and } g(t) \geq m|t - t_{2i}| \text{ on } [(t_{2i-1} + t_{2i})/2, t_{2i}].,$$

i.e. $g(t) \geq m\ dist(t, \partial A^*)$ on A_i^*. □

In order to reconsider the second line of (2.4), we need more information about H_u which vanishes identically on I^*. The following lemma addresses a class of problems where the objective function contains a quadratic control term as it is the case in many applications.

LEMMA 5.4. *Suppose that the objective function contains a quadratic control term*

$$L(x,u) = \bar{L}(x,u) + \frac{\alpha}{2}u^2,$$

with some $\alpha > 0$. *If the function* $f_u(x^*,u^*)^T p^* + \bar{L}_u(x^*,u^*)$ *has a nonzero slope when entering and leaving the active set, then also the second line of* (2.4) *is true.*

The form of L implies $H_u = f_u^T p + \bar{L}_u + \alpha u^*$ and

$$u^* = -\frac{1}{\alpha}(f_u(x^*,u^*)^T p^* + \bar{L}_u(x^*,u^*)) \text{ on } I^*.$$

The proof of this lemma is similar to that one given in Lemma 5.3.

6. Implementation. In this section we want to touch upon some of the details of the implementation of the algorithm. The unconstrained version of the algorithm presented in the previous sections can be given as the following system of nonlinear equations

$$F(z) = F(p,x,u) = \begin{pmatrix} \dot{x} - f(x,u,t) \\ \dot{p} + H_x(p,x,u) \\ H_u(p,x,u) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{6.1}$$

for $z = (p,x,u)^T$ which satisfy the boundary conditions $x(0) = x_0$ and $p(T) = 0$.

The Fréchet -derivative of F is given by

$$F'(p,x,u) \begin{pmatrix} \pi \\ \xi \\ \nu \end{pmatrix} = \begin{pmatrix} 0 & D - f_x & -f_u \\ D + f_x^T & H_{xx} & H_{xu} \\ f_u^T & H_{ux} & H_{uu} \end{pmatrix} \begin{pmatrix} \pi \\ \xi \\ \nu \end{pmatrix} \tag{6.2}$$

with $D = \frac{d}{dt}$ and all other components as multiplication operators.

If one considers the constrained optimal control problem with the corresponding system of nonlinear equations $\mathcal{F}(z) = 0$, then the projected Newton step requires a decision on the set of active indices. In the implementation we used the rule

$$\bar{A}(z) = \{t \in [0,T] : |H_u(z)(t)| \geq \|.5\mathcal{F}(z)\|_Y^{\bar{p}}\}.$$

Then the projected Newton step is computed by solving for (π, ξ, ν) with $\pi(T) = \xi(0) = 0$

$$\begin{aligned} Q_I F'(p,x,u) Q_I \begin{pmatrix} \pi \\ \xi \\ \nu \end{pmatrix} &= \begin{pmatrix} 0 & D - f_x & -f_u\chi_I \\ D + f_x^T & H_{xx} & H_{xu}\chi_I \\ \chi_I f_u^T & \chi_I H_{ux} & \chi_I H_{uu}\chi_I \end{pmatrix} \begin{pmatrix} \pi \\ \xi \\ \nu \end{pmatrix} \\ = -Q_I F(p,x,u) &= \begin{pmatrix} -\dot{x} + f(x,u,t) \\ -\dot{p} - f_x^T p - L_x \\ -\chi_I(f_u^T p + L_u) \end{pmatrix} \end{aligned} \tag{6.3}$$

where

$$I = \{t \in [0,T] : |H_u(t)| < \|\mathcal{F}(z)\|_Y^{\bar{p}}\}.$$

The new control is then computed as

$$u_+ = \mathcal{P}(u + \chi_I \nu - \chi_A H_u).$$

The intermediate new state and adjoint variable are given by

$$(p_{1/2},\ x_{1/2}) = (p,\ x) + (\pi,\ \xi).$$

If one observes that in (6.3) π and ξ appear with derivatives, we can derive a differential equation for the sum $x_{1/2} = x + \xi$ and likewise for $p_{1/2}$. Hence, a more efficient way to compute $(p_{1/2}, x_{1/2})$ is to solve the following differential equations (all unsubscripted quantities are evaluated along the current values (p_c, x_c, u_c)) :

$$\begin{aligned} \dot{x}_{1/2} - f_x x_{1/2} &= -f_x x + f + f_u \chi_I \nu, \\ \dot{p}_{1/2} + f_x^T p_{1/2} + H_{xx} x_{1/2} &= H_{xx} x - L_x - H_{xu}\chi_I \nu. \end{aligned} \tag{6.4}$$

with boundary conditions $x_{1/2}(0) = x_0, p_{1/2}(T) = 0$. In addition, if we use the invertibility of $\chi_I H_{uu}\chi_I$ for all $t \in I$, then ν can be expressed in terms of $x_{1/2}, p_{1/2}, x, p$:

$$\chi_I \nu = -(H^I_{uu})^+ (f_u^T p_{1/2} + H_{ux}(x_{1/2} - x) + L_u),$$

where we substitute

$$(H^I_{uu})^+ = \begin{cases} (\chi_I H_{uu}\chi_I)^{-1}(t) & \text{if} \quad (\chi_I H_{uu}\chi_I)(t) \neq 0 \\ 0 & \text{if} \quad (\chi_I H_{uu}\chi_I)(t) = 0 \end{cases}. \tag{6.5}$$

Hence a linear two point boundary value problem needs to be solved at each iteration: Solve for $p_{1/2}, x_{1/2}$ with $p_{1/2}(T) = 0, x_{1/2}(0) = 0$

$$\begin{aligned} \dot{x}_{1/2} &+ (-f_x + f_u(H^I_{uu})^+ H_{ux})x_{1/2} + f_u(H^I_{uu})^+ f_u^T p_{1/2} \\ &= (-f_x + f_u(H^I_{uu})^+ H_{ux})x + f - f_u(H^I_{uu})^+ L_u, \\ \dot{p}_{1/2} &+ (f_x^T - H_{xu}(H^I_{uu})^+ f_u^T)p_{1/2} + (H_{xx} - H_{xu}(H^I_{uu})^+ H_{ux})x_{1/2} \\ &= (H_{xx} - H_{xu}(H^I_{uu})^+ H_{ux})x - L_x + H_{xu}(H^I_{uu})^+ L_u. \end{aligned} \tag{6.6}$$

At the end of each iteration a smoothing step has to be carried out.

$$\begin{aligned} x_+(t) &= \int_0^t f(x_{1/2}(s), u_+(s), s)\ ds + x_0 \\ p_+(t) &= \int_t^T H_x(p_{1/2}(s), x_{1/2}(s), u_+(s))\ ds \end{aligned} \tag{6.7}$$

Termination and the identification of the active set is based on the size of the residual which can be computed as follows:

$$F(z_+) = \begin{pmatrix} \dot{x}_+ - f(x_+, u_+, t) \\ \dot{p}_+ + H_x(p_+, x_+, u_+) \\ u_+ - \mathcal{P}(u_+ - H_u(p_+, x_+, u_+)) \end{pmatrix} = \begin{pmatrix} f(x_{1/2}, u_+, t) - f(x_+, u_+, t) \\ H_x(p_+, x_+, u_+) - H_x(p_{1/2}, x_{1/2}, u_+) \\ u_+ - \mathcal{P}(u_+ - H_u(p_+, x_+, u_+)) \end{pmatrix} \tag{6.8}$$

We also list the size of the step $\|z_+ - z_c\|_X$ which we calculate by (the intermediate iterate $x_{1/2}$ has a corresponding iterate in the previous step denoted by $x_{-1/2}$)

$$\begin{aligned} \|z_+ - z_c\|_X &= \max \{\|x_+ - x_c\|_\infty + \|\dot{x}_+ - \dot{x}_c\|_\infty, \|p_+ - p_c\|_\infty + \|\dot{p}_+ - \dot{p}_c\|_\infty, \|u_+ - u_c\|_\infty\} \\ &= \max \{\|x_+ - x_c\|_\infty + \|f(x_{1/2}, u_+, \cdot) - f(x_{-1/2}, u_c, \cdot)\|_\infty, \\ &\qquad \|p_+ - p_c\|_\infty + \|H_x(p_{1/2}, x_{1/2}, u_+) - H_x(p_{-1/2}, x_{-1/2}, u_c)\|_\infty, \|u_+ - u_c\|_\infty\} \end{aligned} \tag{6.9}$$

We use the following example to illustrate the results: Let

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} (1 - x_2^2)x_1 - x_2 + u \\ x_1 \end{pmatrix}, \quad L(x, u) = \frac{1}{2}(x_1^2 + x_2^2 + u^2)$$

and $T = 3, x(0) = (0, 1)^T$. Furthermore let

$$0 \leq u(t) \leq 0.8,$$

and starting data $x_0 \equiv (0,1)^T, p_0 \equiv (0,0)^T, u_0 \equiv 0$.

In Tables 6.1, 6.2, and 6.3 we tabulate, for different values of $\bar{p}$ the progress of the iteration for the example above. The two point boundary value problem (6.4) was solved with the trapezoid rule extrapolation approach used in [11] and the integration in (4.3) was done with the trapezoid rule. A uniform mesh of 1400 points was used.

For each iterate k we tabulate the norm of the nonlinear residual

$$\rho = \|(\dot{x} - f, \dot{p} + H_x, u - \mathcal{P}(u - H_u))\|_\infty$$

the ratio $\rho_{k+1}/\rho_k^{\bar{p}}$, and the norm of the step $\sigma = \|(p_{k+1} - p_k, x_{k+1} - x_k, u_{k+1} - u_k)\|_\infty$.

We see from the tables that the convergence follows the predictions of Theorem 4.1 until the residual can be reduced no further as a result of truncation error effects. In several numerical experiments with different numbers of mesh points, a sharp increase in $\rho_{k+1}/\rho_k^{\bar{p}}$ seemed to be an indicator that truncation error effects were dominating the computation.

TABLE 6.1
$\bar{p} = 0.6$

k	ρ_k	ρ_k/ρ_{k-1}	$\rho_k/\rho_{k-1}^{\bar{p}}$	σ_k
1	0.8669D+00	0.867	0.867	0.3308D+01
2	0.1746D+00	0.201	0.207	0.1331D+01
3	0.1054D+00	0.604	0.856	0.1656D+00
4	0.7004D-01	0.664	1.042	0.1706D+00
5	0.3290D-01	0.470	0.799	0.1463D+00
6	0.9885D-02	0.300	0.595	0.8485D-01
7	0.1604D-02	0.162	0.409	0.2830D-01
8	0.1328D-03	0.083	0.300	0.4816D-02
9	0.3456D-05	0.026	0.155	0.4380D-03

TABLE 6.2
$\bar{p} = 0.75$

k	ρ_k	ρ_k/ρ_{k-1}	$\rho_k/\rho_{k-1}^{\bar{p}}$	σ_k
1	0.8669D+00	0.867	0.867	0.2135D+01
2	0.1746D+00	0.201	0.216	0.1331D+01
3	0.7894D-01	0.452	1.083	0.2370D+00
4	0.3058D-01	0.387	1.379	0.2007D+00
5	0.5309D-02	0.174	0.993	0.9241D-01
6	0.2431D-03	0.046	0.629	0.1671D-01
7	0.1638D-05	0.007	0.432	0.7524D-03
8	0.1691D-05	1.033	806.927	0.9862D-06
9	0.1691D-05	1.000	769.034	0.5488D-12

TABLE 6.3
$\bar{p} = 0.9$

k	ρ_k	ρ_k/ρ_{k-1}	$\rho_k/\rho_{k-1}^{\bar{p}}$	σ_k
1	0.8669D+00	0.867	0.867	0.2135D+01
2	0.1746D+00	0.201	0.226	0.1331D+01
3	0.6306D-01	0.361	1.460	0.3048D+00
4	0.1512D-01	0.240	2.187	0.1895D+00
5	0.9063D-03	0.060	1.715	0.4991D-01
6	0.3483D-05	0.004	1.044	0.2929D-02
7	0.1691D-05	0.485	11287.238	0.9399D-05
8	0.1691D-05	1.000	41449.044	0.3974D-11
9	0.1691D-05	1.000	41448.895	0.8188D-13

With Table 6.4 we want to illustrate the effect of Assumption 2.3 on the rate of convergence. If this assumption does not hold we are no longer guaranteed a local superlinear rate of convergence as in Theorem 4.1. In Lemma 5.4 we give a sufficient condition for Assumption 2.3 to be true. For the example under consideration we have

$$f_u(x^*, u^*)^T p^* + \bar{L}_u(x^*, u^*) = p_1^*$$

which looks like a parabola $-(t-2)^2 + 1$ with negative curvature. Here we impose only upper bounds on the controls. If the upper bound is relatively high, like 0.95, the slope of p_1^* at the boundary of the active set is small. The assumption in Lemma 5.4 is still satisfied but we can see a slower rate of convergence locally compared to an example where the upper bound on the control is set to 0.8 yielding a steeper slope of p_1^* at the boundary of the active set.

We have 1400 discretization points and select $\bar{p} = 0.6$.

TABLE 6.4
$\bar{p} = 0.6$

	u_{max}=0.95		u_{max}=0.8	
k	ρ_k	ρ_k/ρ_{k-1}	ρ_k	ρ_k/ρ_{k-1}
1	0.9107D+00		0.9107D+00	
2	0.1799D+00	0.198	0.2503D+00	0.275
3	0.6510D-02	0.036	0.1936D+00	0.774
4	0.4760D-02	0.731	0.1215D+00	0.628
5	0.2568D-02	0.540	0.5380D-01	0.443
6	0.9994D-03	0.389	0.1357D-01	0.252
7	0.2761D-03	0.276	0.1707D-02	0.126
8	0.5716D-04	0.207	0.1279D-03	0.075
9	0.9666D-05	0.169	0.8496D-05	0.066
10	0.2229D-05	0.231	0.4072D-05	0.479

Acknowledgments. The authors appreciate the work of the two very careful referees.

REFERENCES

[1] D. B. BERTSEKAS, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Autom. Control, 21 (1976), pp. 174–184.

[2] ———, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.

[3] J. W. Daniel, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, 1971.
[4] V. F. Demyanov and A. M. Rubinov, *Approximate Methods in Optimization Theory*, Elsevier, New York, 1970.
[5] J. C. Dunn, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, SIAM J. Control Optim., 17 (1979), pp. 187–211.
[6] ———, *Newton's method and the Goldstein step-length rule for constrained minimization problems*, SIAM J. Control Optim., 16 (1980), pp. 659–674.
[7] ———, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Control Optim., 19 (1981), pp. 368–400.
[8] J. C. Dunn and T. Tian, *On the gradient projection method for optimal control problems with nonnegative L2 inputs*. SIAM J. Control and Optimization, to appear.
[9] ———, *Variants of the Kuhn-Tucker sufficient conditions in cones of nonnegative functions*. SIAM J. Control and Optimization, to appear.
[10] A. A. Goldstein, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.
[11] C. T. Kelley and E. W. Sachs, *A pointwise quasi-Newton method for unconstrained optimal control problems*, Numer. Math., 55 (1989), pp. 159–176.
[12] ———, *Mesh independence of the gradient projection method for optimal control problems*, SIAM J. Control and Optimization, 30 (1992), pp. 477–493.
[13] ———, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. on Sci. Comp., 15 (1994), pp. 645–667.
[14] E. S. Levitin and B. T. Polyak, *Constrained optimization methods*, USSR Comput. Math. Phys., 6 (1966), pp. 1–50.
[15] H. Maurer, *The two-norm approach for second order sufficiency conditions in mathematical programming and optimal control*, Tech. Report 6/92 - N, Inst. f. Angew. Mathem. und Informatik, Universität Münster, 1992.
[16] D. Orrell and V. Zeidan, *Another Jacobi sufficiency criterion for optimal control with smooth constraints*, J. Optim. Theory Appl., 58 (1988), pp. 282–300.
[17] E. W. Sachs, *Convergence of algorithms for perturbed optimization problems*, Annals of Operations Research, 27 (1990), pp. 311–342.