CONSISTENT APPROXIMATIONS FOR OPTIMAL CONTROL PROBLEMS BASED ON RUNGE-KUTTA INTEGRATION

by

A. Schwartz and E. Polak

Memorandum No. UCB/ERL M94/21

10 April 1994

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

CONSISTENT APPROXIMATIONS FOR OPTIMAL CONTROL PROBLEMS BASED ON RUNGE-KUTTA INTEGRATION[†]

by

A. Schwartz and E. Polak
Department of Electrical Engineering
and Computer Sciences
University of California
Berkeley, CA 94720, U.S.A.

ABSTRACT

This paper explores the use of Runge-Kutta integration methods in the construction of families of finite dimensional, consistent approximations to non-smooth, control and state constrained optimal control problems. Consistency is defined in terms of epiconvergence of the approximating problems and hypoconvergence of their optimality functions. A major consequence of this concept of consistency is that stationary points and global solutions of the approximating discrete time optimal control problems can only converge to stationary points and global solutions of the original optimal control problem. The construction of consistent approximations required the introduction of appropriate finite dimensional subspaces of the space of controls and the extension of the standard Runge-Kutta methods to piecewise continuous functions.

It is shown that unless a non-Euclidean inner product and norm are used on the control space, in solving the approximating discrete time optimal control problems, considerable ill-conditioning may result.

Key words. optimal control, discretization theory, consistent approximations, runge-kutta integration

AMS subject classifications. 49J15, 49M25, 49J45, 65L06

[†] The research reported herein was sponsored by the Air Force Office of Scientific Research contract F49620-93-1-0165 and the National Science Foundation grant ECS-93-02926.

1. INTRODUCTION.

Except for very special cases, optimal control problems can only be solved numerically, using such discretization techniques as numerical integration (see, *e.g.* [6,10,11,19]) or collocation (see, *e.g.*, [12,21,26,29,32]). Numerical integration is used in two ways: to implement conceptual optimal control algorithms (see, e.g., [15,30]), and to construct approximating discrete time optimal control problems that can then be solved by any applicable discrete time optimal control or nonlinear programming algorithm. In this paper we are concerned with the latter. With a few exceptions, such as [16,31], most authors, for example [6,10,19,20], dealing with the construction of approximating discrete time optimal control problems, assume that Euler's method is used for integration.

The central question in discretization theory is whether solutions to the approximating problems converge to solutions of the original problem. In the context of optimization, the term ''solution'' is used ambiguously; it can mean ''global solution'', ''local solution'', or ''stationary point''. Convergence of global solutions, or in some cases, of stationary points of the approximating problems to those of the original problem, was treated in [6,10,11,13,19,20,24]. Rate of convergence of stationary points of approximating problems to those of the original problem was explored in [16] for a class of unconstrained problems. Possibly the most extensive treatment of the the issues of approximation of general, nonsmooth, constrained optimal control problems by approximating problems obtained by Euler's method, can be found in [19]. In particular, we find in [19] proofs of the existence of solutions of the approximating problems and convergence of discrete controls, satisfying an approximate discrete time Maximum Principle, to a control satisfying the Maximum Principle for the original problem.

Daniel [13] presents one of the first attempts to characterize consistency of approximations to an optimization problem, and establishes conditions for the convergence of approximate global solutions to approximating problems, obtained by discretization, to global solutions of the original problem. The more recent and more elegant epiconvergence theory in [2,14], is set within the framework of a general theory of convergence of set valued maps and yields the same results in a simpler, more straightforward manner. Neither theory addresses issues of computation.

The theory of consistent approximations, presented in [24], is directed towards the construction of finite dimensional approximating problems that can be used in conjunction with diagonalization strategies and nonlinear programming algorithms to efficiently obtain an approximate, numerical ''solution'' to an original infinite dimensional problem. The theory in [24] considers pairs consisting of an abstract optimization problem $\mathbf{P}$, defined on a normed space $\mathcal{H}$, and of an *optimality function* $\theta(\cdot)$, whose zeros are the stationary points of $\mathbf{P}$. These pairs are approximated by pairs $\mathbf{P}_N$ and $\theta_N(\cdot)$, $N = 1,2,3,\ldots,$ that are defined on nested finite dimensional subspaces $\mathcal{H}_N$ of $\mathcal{H}$ with $\theta_N(\cdot)$ an optimality function for $\mathbf{P}_N$ (rather than a discretization of $\theta(\cdot)$). Consistency of approximation is characterized in terms of epiconvergence of the $\mathbf{P}_N$ to $\mathbf{P}$ and (in the simplest case) of hypoconvergence of the $\theta_N(\cdot)$ to $\theta(\cdot)$.

Epiconvergence of the approximating problems ensures convergence of global minimizers (and uniformly strict local minimizers) of the approximating problems to global minimizers (local minimizers) of the original problem. Hypoconvergence of optimality functions ensures, directly or indirectly, several desirable properties: *(i)* stationary points of the approximating problems converge to stationary points of the original problem; *(ii)* the mathematical characterization of the constraints of the approximating problems must satisfy certain consistency conditions; and *(iii)* derivatives of the cost and constraint functions of the approximating problems converge to those of the original problem.

In [24], consistent approximations were constructed for control and state inequality constrained optimal control problems using Euler's method and control subspaces spanned by piecewise constant functions. In this paper we show that a large class of higher order, explicit Runge-Kutta (RK) methods can be used to construct consistent approximations for the same problems. Two issues had to be addressed: the selection of the finite dimensional control subspaces and the isometric transformation of the resulting approximating optimal control problems into mathematical programming problems. The selection of the control subspaces for use with RK methods is significantly more complex than for Euler's method and affects the precision of integration accuracy as well as original problem solution approximation. Isometric transformations of the approximating optimal control problems into mathematical problems defined on a Euclidean space preserve problem conditioning. As demonstrated by our computational results in Section 6, a considerable deterioration of conditioning can take when natural, but non-isometric transformations are used.

This paper is organized as follows. Section 2 summarizes the relevant aspects of the theory of consistent approximations. Section 3 introduces the optimal control problem and develops an optimality function for it. In section 4 the approximating problems are constructed, by defining appropriate finite-dimensional control spaces and constraint sets, and by defining approximate cost functions using an extension of RK integration methods. Epiconvergence is proved. In section 5, the optimality functions for the approximating problems are derived. These are shown to hypoconverge to the optimality function for the original problem. Hence, the approximating problems are shown to be consistent approximations for the original problem. Section 6 presents some numerical results.

2. CONSISTENT APPROXIMATIONS

In [24] we find a theory of consistent approximations to an abstract optimization problem, defined on a normed space. The theory uses two concepts: epiconvergence of the epigraphs of of the approximating cost functions on the approximating constraint sets, which results in a type of "zero order" approximation, and the satisfaction of half of the relations that ensure the hypoconvergence of the hypographs of optimality functions, of the approximating problems, which ensures a type of "first order" approximation. We will briefly review those results. Let $\mathcal{H}$ be a normed linear space, with norm $\| \cdot \|$, let $\mathbf{B} \subset \mathcal{H}$ be a closed, convex set and consider the problem

$$\mathbf{P} \qquad \min_{\eta \in \mathbf{H}} \psi(\eta) \tag{2.1a}$$

where $\psi : \mathbf{B} \to \mathbb{R}$ is (at least) lower semi-continuous, and $\mathbf{H} \subset \mathbf{B}$ is the constraint set. Next, let $\mathbb{N} \triangleq \{1,2,3,\ldots\}$, let $\mathbf{N}$ be an infinite subset of $\mathbb{N}$, and let $\{\mathcal{H}_N\}_{N \in \mathbf{N}}$ be a family of finite dimensional subspaces of $\mathcal{H}$ such that $\mathcal{H}_N = \mathcal{H}$ if $\mathcal{H}$ is finite dimensional ($\mathbb{R}^n$) and $\mathcal{H}_{N_1} \subset \mathcal{H}_{N_2}$, for all $N_1 N_2 \in \mathbf{N}$ such that $N_1 < N_2$, otherwise. Now consider the family of approximating problems

$$\mathbf{P}_N \qquad \min_{\eta \in \mathbf{H}_N} \psi_N(\eta), \quad N \in \mathbf{N}, \tag{2.1b}$$

where $\psi_N : \mathcal{H}_N \to \mathbb{R}$ is (at least) lower semicontinuous, and $\mathbf{H}_N \subset \mathcal{H}_N$.

Definition 2.1. We will say that the problems in the family $\{\mathbf{P}_N\}_{N \in \mathbf{N}}$ *converge epigraphically* to $\mathbf{P}$ ($\mathbf{P}_N \to^{Epi} \mathbf{P}$) if

(a) for every $\eta \in \mathbf{H}$, there exists a sequence $\{\eta_N\}_{N \in \mathbf{N}}$, with $\eta_N \in \mathbf{H}_N$, such that $\eta_N \to \eta$ and $\overline{\lim}\, \psi_N(\eta_N) \le \psi(\eta)$;

(b) for every infinite sequence $\{\eta_N\}_{N \in K}$, where $K \subset \mathbf{N}$, satisfying $\eta_N \in \mathbf{H}_N$, for all $N \in K$ and $\eta_N \to^K \eta$, we have that $\eta \in \mathbf{H}$ and $\underline{\lim}\, \psi_N(\eta_N) \ge \psi(\eta)$. □

Epiconvergence does not require derivative information for its characterization. Hence we view epiconvergence as "zero order" approximation property. In [2,14,24] we find the following result:

Theorem 2.2. Suppose that $\mathbf{P}_N \to^{Epi} \mathbf{P}$. *(a)* If $\{\hat{\eta}_N\}_{N \in \mathbf{N}}$ is a sequence of global minimizers of the $\mathbf{P}_N$, and $\hat{\eta}$ is any accumulation point of $\{\hat{\eta}_N\}_{N \in \mathbf{N}}$, then $\hat{\eta}$ is a global minimizer of $\mathbf{P}$. *(b)* If $\{\hat{\eta}_N\}_{N \in \mathbf{N}}$ is a sequence of strict local minimizers of the $\mathbf{P}_N$, whose radii of attraction are bounded away from zero, and $\hat{\eta}$ is any accumulation point of $\{\hat{\eta}_N\}_{N \in \mathbf{N}}$, then $\hat{\eta}$ is a local minimizer of $\mathbf{P}$. □

Epigraphical convergence does not eliminate the possibility of stationary points of $\mathbf{P}_N$, converging to a non-stationary point of $\mathbf{P}$: a most annoying result from a numerical optimization point of view. For example, let $\mathcal{H} = \mathbb{R}^2$ with $\eta = (x, y)$, and let $f(\eta) = f_N(\eta) = (x-2)^2$, $N \in \mathbb{N}$. Choose

$$\mathbf{H} \triangleq \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 - 2 \le 0\}, \tag{2.2a}$$

$$\mathbf{H}_N \triangleq \{(x,y) \in \mathbb{R}^2 \mid (x-y)^2(x^2+y^2-2) \le 0,\ x^2+y^2 \le 2 + 1/N\},\ N \in \mathbb{N}. \tag{2.2b}$$

Then we see that $\mathbf{P}_N \to^{Epi} \mathbf{P}$. Nevertheless, the point (1,1) is feasible and satisfies the F. John optimality condition for all $\mathbf{P}_N$, but it is not a stationary point for the problem $\mathbf{P}$. The reason for this is an incompatibility of the constraint sets $\mathbf{H}_N$ with the constraint set $\mathbf{H}$ which shows up only at the level of optimality conditions. To eliminate the possibility of this happening, at least for first order non-stationary points, optimality functions were introduced in [24] as a tool for ensuring a kind of "first order" approximation result, which, implicitly, enforces convergence of derivatives, and restricts the forms chosen for the description of the sets $\mathbf{H}$ and $\mathbf{H}_N$.

Definition 2.4. We will say that a function $\theta : \mathbf{B} \to \mathbb{R}$ is an *optimality function* for $\mathbf{P}$ if *(i)* $\theta(\cdot)$ is (at least) upper semi-continuous, *(ii)* $\theta(\eta) \le 0$ for all $\eta \in \mathbf{H}$, and *(iii)* for $\hat{\eta} \in \mathbf{H}$, $\theta(\hat{\eta}) = 0$ if $\hat{\eta}$ is a local minimizer for $\mathbf{P}$. Similarly, we will say that a function $\theta_N : H_N \to \mathbb{R}$ is an *optimality function* for $\mathbf{P}_N$ if *(i)* $\theta_N(\cdot)$ is (at least) upper semi-continuous, *(ii)* $\theta_N(\eta) \le 0$ for all $\eta \in \mathbf{H}_N$, and *(iii)* if $\hat{\eta}_N \in \mathbf{H}_N$ is a local minimizer for $\mathbf{P}_N$ then $\theta_N(\hat{\eta}_N) = 0$ □

Definition 2.5. Consider the problems $\mathbf{P}, \mathbf{P}_N$, defined in (2.1a,b). Let $\theta(\cdot), \theta_N(\cdot), N \in \mathbf{N}$, be optimality functions for $\mathbf{P}$, $\mathbf{P}_N$, respectively. We will say that the pairs $(\mathbf{P}_N, \theta_N)$, in the sequence $\{ (\mathbf{P}_N, \theta_N) \}_{N=1}^{\infty}$ are *consistent approximations* to the pair $(\mathbf{P}, \theta)$, if *(i)* $\mathbf{P}_N \to^{Epi} \mathbf{P}$, and *(ii)* for any sequence $\{ \eta_N \}_{N \in K}$, $K \subset \mathbf{N}$, with $\eta_N \in \mathbf{H}_N$ for all $N \in K$, such that $\eta_N \to \eta$, $\overline{\lim}\, \theta_N(\eta_N) \le \theta(\eta)$. □

Note that the last part of this definition, concerning convergence of the optimality functions, rules out the possibility of stationary points (points such that $\theta_N(\eta_N) = 0$) for the approximating problems converging to non-stationary points of the original problem. In the sequel, we will prove a stronger condition, namely convergence of the hypographs of $\theta_N(\cdot)$ to the hypograph of $\theta(\cdot)$, than is required by Definition 2.5 (that is $-\theta_N(\cdot) \to^{Epi} -\theta(\cdot)$).

3. PROBLEM DEFINITION

We will consider optimal control problems with dynamic systems described by ordinary differential equations,

$$\dot{x}(t) = h(x(t), u(t)), \text{ a.e. for } t \in [0,1], \ x(0) = x_0, \tag{3.1}$$

where $x(t) \in \mathbb{R}^{n_x}$, $u(t) \in \mathbb{R}^m$, and hence $h : \mathbb{R}^{n_x} \times \mathbb{R}^m \to \mathbb{R}^{n_x}$.

To establish continuity and differentiability of solutions of (3.1) with respect to controls, one must assume that the controls are bounded in $L_\infty^m[0,1]$. However, our approximating subspaces are dense in $L_2^m[0,1]$, but not in $L_\infty^m[0,1]$. Since it does not appear to be possible to establish differentiability of solutions of (3.1) with respect to controls in $L_2^m[0,1]$, we will, as in [23], assume that the controls are elements of the pre-Hilbert space

$$L_{\infty,2}^m[0,1] \triangleq (L_\infty^m[0,1], \langle \cdot, \cdot \rangle_2, \|\cdot\|_2), \tag{3.2a}$$

which consists of the elements of $L_\infty^m[0,1]$, but is endowed with the $L_2^m[0,1]$ inner product and norm. Note that $L_{\infty,2}^m[0,1]$ is dense in $L_2^m[0,1]$.

Since we also vary initial states, we will work in the pre-Hilbert space

$$H_{\infty,2} \triangleq \mathbb{R}^{n_x} \times L^m_{\infty,2}[0,1] \triangleq (\mathbb{R}^{n_x} \times L^m_\infty[0,1], \langle \cdot,\cdot \rangle_H, \| \cdot \|_H), \tag{3.2b}$$

which is a dense subspace of the Hilbert space

$$H_2 = \mathbb{R}^{n_x} \times L^m_2[0,1]\,. \tag{3.2c}$$

The inner product $\langle \cdot, \cdot \rangle_H$ and norm $\| \cdot \|_H$, on H_2, and hence also on $H_{\infty,2}$, are defined as follows. For any $\eta = (\xi, u) \in H_2$ and $\eta' = (\xi', u') \in H_2$,

$$\langle \eta, \eta' \rangle_H \triangleq \langle \xi, \xi' \rangle + \langle u, u' \rangle_2, \tag{3.2d}$$

where $\langle \xi, \xi' \rangle$ denotes the Euclidean inner product, and the L_2 inner product $\langle u, u' \rangle_2$ is defined by $\langle u, v \rangle_2 \triangleq \int_0^1 \langle u(t), v(t) \rangle \, dt$. Consequently, for any $\eta = (\xi, u) \in H_2$,

$$\| \eta \|_H^2 \triangleq \langle \eta, \eta \rangle_H = \| \xi \|^2 + \| u \|_2^2\,. \tag{3.2e}$$

Next, let $U \subset B(0, \rho_{\max}) \triangleq \{ u \in \mathbb{R}^m \mid \| u \| \leq \rho_{\max} \}$ be a compact convex set with non-empty interior, where $\rho_{\max}$ is sufficiently large to ensure that all the control functions $u(\cdot)$ which we expect to deal with take values in the interior of $B(0, \rho_{\max})$. We define

$$\mathbf{U} \triangleq \{ u \in L^m_{\infty,2}[0,1] \mid u(t) \in U \,,\ \forall t \in [0,1] \} \tag{3.3a}$$

and denote the set of admissible initial state-control pairs, $\eta \triangleq (\xi, u)$, by

$$\mathbf{H} \triangleq \mathbb{R}^{n_x} \times \mathbf{U} \subset H_{\infty,2}\,. \tag{3.3b}$$

The set $\mathbf{H}$ is contained in the larger set

$$\mathbf{B} \triangleq \mathbb{R}^{n_x} \times \{ u \in L^m_{\infty,2}[0,1] \mid u(t) \in B(0, \rho_{\max}) \,,\ \forall t \in [0,1] \} \subset H_{\infty,2}\,. \tag{3.3c}$$

Finally, we will denote solutions of (3.1) corresponding to a particular $\eta \in \mathbf{H}$ by $x^\eta(\cdot)$. We will consider the following canonical minimax optimal control problem:

CP

$$\min_{\eta \in \mathbf{H}} \{ \psi_o(\eta) \mid \psi_c(\eta) \leq 0 \}\,, \tag{3.4a}$$

where the objective function, $\psi_o : \mathbf{B} \to \mathbb{R}$, and the state endpoint constraint function, $\psi_c : \mathbf{H} \to \mathbb{R}$ are defined by

$$\psi_o(\eta) \triangleq \max_{\nu \in \mathbf{q}_o} f^\nu(\eta)\,, \quad \psi_c(\eta) \triangleq \max_{\nu \in \mathbf{q}_c + q_o} f^\nu(\eta)\,, \tag{3.4b}$$

with $f^\nu : \mathbf{H} \to \mathbb{R}$ defined by

$$f^\nu(\eta) = \zeta^\nu(\xi, x^\eta(1))\,, \tag{3.4c}$$

with $\zeta^\nu : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}$, and we have used the notation $\mathbf{q}_o \triangleq \{ 1, 2, \ldots, q_o \}$, $\mathbf{q}_c \triangleq \{ 1, 2, \ldots, q_c \}$

(with q_o and q_c arbitrary integers). The set $\mathbf{q}_c + q_o = \{ 1 + q_o, \ldots, q_c + q_o \}$. In what follows, we will let $\mathbf{q} \triangleq \mathbf{q}_o \cup \{ \mathbf{q}_c + q_o \}$. If we define $\mathbf{H}' \triangleq \{ \eta \in \mathbf{H} \mid \psi_c(\eta) \le 0 \}$, we see that **CP**, with **H** replaced by **H**′, is of the form of the problem **P** in (2.1a).

Various optimal control problems, such as non-autonomous, integral cost, and free-time problems, can be transcribed into this canonical form. Also, the endpoint constraint in (3.4a) can be discarded by setting $\psi_c(\eta) \equiv -\infty$, and control unconstrained problems can be included by choosing $\rho_{\max}$ and U sufficiently large to ensure that the solutions $u^*(\cdot)$ of **CP** take values in the interior of U.

Properties of the Defining Functions. We will require the following assumptions:

Assumption 3.1.

(a) The function $h(\cdot,\cdot)$ in (3.1) is continuously differentiable, and there exists a Lipschitz constant $\kappa \in [1,\infty)$ such that for all $x', x'' \in \mathbb{R}^{n_x}$, and $v', v'' \in B(0, \rho_{\max})$ the following relations hold:

$$\| h(x', v') - h(x'', v'') \| \le \kappa [\| x' - x'' \| + \| v' - v'' \|] , \qquad (3.5a)$$

$$\| h_x(x', v') - h_x(x'', v'') \| \le \kappa [\| x' - x'' \| + \| v' - v'' \|] , \qquad (3.5b)$$

$$\| h_u(x', v') - h_u(x'', v'') \| \le \kappa [\| x' - x'' \| + \| v' - v'' \|] , \qquad (3.5c)$$

(b) The functions $\zeta^\nu(\cdot,\cdot)$, $\zeta^\nu_\xi(\cdot,\cdot)$ and $\zeta^\nu_x(\cdot,\cdot)$, with $\nu \in \mathbf{q}$, are Lipschitz continuous on bounded sets. □

The following results can be found in [3]:

Theorem 3.2. If Assumption 3.1 is satisfied then

(i) there exists an $\kappa \in [0,\infty)$ such that for all $\eta', \eta'' \in \mathbf{B}$ and for all $t \in [0,1]$

$$\| x^{\eta'}(t) - x^{\eta''}(t) \| \le \kappa \| \eta' - \eta'' \|_H ;$$

(ii) there exists a $\kappa \in [0,\infty)$ such that for all $\eta \in \mathbf{B}$ and all $t \in [0,1]$

$$\| x^{\eta}(t) \| \le \kappa (1 + \| \xi \|) ;$$

(iii) the functions $\psi_o : \mathbf{B} \to \mathbb{R}$ and $\psi_c : \mathbf{B} \to \mathbb{R}$ are Lipschitz continuous on bounded sets;

(iv) the functions $f^\nu(\cdot)$, $\nu \in \mathbf{q}$, have continuous Gateaux differentials $Df^\nu : \mathbf{B} \times H_{\infty,2} \to \mathbb{R}^{n_x}$ that have the form $Df^\nu(\eta ; \delta\eta) = \langle \nabla f^\nu(\eta), \delta\eta \rangle_H$;

(v) the gradients $\nabla f^\nu(\eta) = (\nabla_\xi f^\nu(\eta), \nabla_u f^\nu(\eta)) \in H_{\infty,2}$, $\nu \in \mathbf{q}$, are given by

$$\nabla_\xi f^\nu(\eta) = \nabla_\xi \zeta^\nu(\xi, x^\eta(1)) + p^{\eta,\nu}(0) , \qquad (3.6a)$$

$$\nabla_u f^\nu(\eta)(t) = h_u(x(t), u(t))^T p^{\eta,\nu}(t) , \quad \forall t \in [0,1], \qquad (3.6b)$$

where $p^{\eta,\nu}(t) \in \mathbb{R}^{n_x}$ is the solution to the adjoint equation

$$\dot{p}^{\nu} = -h_x(x,u)^T p^{\nu} \ , \quad p^{\nu}(1) = \nabla_x \zeta^{\nu}(\xi, x(1)) \, , \tag{3.6c}$$

and are Lipschitz continuous on bounded sets in **B**.

(vi) for any $\nu \in \mathbf{q}$, and any $\eta \in \mathbf{B}$ and $\delta\eta \in H_{\infty,2}$, the directional derivative exists and is given by

$$df^{\nu}(\eta\,;\delta\eta) = Df^{\nu}(\eta\,;\delta\eta) = \langle \nabla f^{\nu}(\eta), \delta\eta \rangle_H \, , \tag{3.6d}$$

furthermore, it is Lipschitz continuous on bounded sets in both arguments. □

Optimality Conditions. Referring to [9], the following result holds because of Theorem 3.2:

Theorem 3.3 For any $\eta \in \mathbf{B}$, let

$$\psi_c(\eta)_+ \triangleq \max\{\, 0, \psi_c(\eta) \,\} \ , \tag{3.7a}$$

and for any $\eta, \eta' \in \mathbf{B}$ and $\sigma > 0$, let

$$\Psi(\eta, \eta') \triangleq \max\{\, \psi_o(\eta) - \psi_o(\eta') - \sigma\psi_c(\eta')_+ \, , \psi_c(\eta) - \psi_c(\eta')_+ \,\} \, . \tag{3.7b}$$

If Assumption 3.1 is satisfied and $\hat{\eta} \in \mathbf{H}$ is a local minimizer of the problem **CP**, then

$$d\Psi_{\eta'}(\hat{\eta}, \hat{\eta}\,; \eta - \hat{\eta}) \geq 0 \, , \quad \forall \eta \in \mathbf{H} \, . \tag{3.8}$$

□

Next we define the optimality function $\theta : \mathbf{B} \to \mathbb{R}$, for **CP**. For any $\eta, \eta' \in \mathbf{H}$ and $\nu \in \mathbf{q}$, we define the first-order quadratic approximation to $f^{\nu}(\cdot)$ at η by

$$\tilde{f}^{\nu}(\eta, \eta') \triangleq f^{\nu}(\eta) + \langle \nabla f^{\nu}(\eta), \eta' - \eta \rangle_H + \tfrac{1}{2}\|\eta' - \eta\|_H^2 \, . \tag{3.9a}$$

Then, the optimality function, with $\sigma > 0$ as in (3.7b), is defined as

$$\theta(\eta) \triangleq \min_{\eta' \in \mathbf{H}} \max\left\{ \max_{\nu \in \mathbf{q}_o} \tilde{f}^{\nu}(\eta, \eta') - \psi_o(\eta) - \sigma\psi_c(\eta)_+ \, , \max_{\nu \in \mathbf{q}_c + q_o} \tilde{f}^{\nu}(\eta, \eta') - \psi_c(\eta)_+ \right\} \, . \tag{3.9b}$$

The existence of the minimum in (3.9b) follows from the convexity of the constraint set **H** and of the max functions in (3.9b) with respect to η', and $\tilde{f}^{\nu}(\eta, \eta') \to \infty$ as $\|\eta'\| \to \infty$ [5, Corollary III.20, p.4 6]. Note that if $f^{\nu}(\eta) \equiv -\infty$ for all $\nu \in \mathbf{q}_c + q_o$, so that $\psi_c(\eta) \equiv -\infty$, then (3.9b) reduces to

$$\theta(\eta) \triangleq \min_{\eta' \in \mathbf{H}} \max_{\nu \in \mathbf{q}_o} f_{\nu}(\eta) + \langle \nabla f^{\nu}(\eta), \eta' - \eta \rangle_H + \tfrac{1}{2}\|\eta' - \eta\|_H^2 - \psi_o(\eta) \, . \tag{3.9c}$$

Referring once again to [3], we have the following result:

Theorem 3.5. Let $\theta : \mathbf{B} \to \mathbb{R}$ be defined by equation (3.9b). If Assumption 3.1 holds then, *(i)* $\theta(\cdot)$ is negative valued and continuous; *(ii)* the relation (3.8) holds if and only if $\theta(\hat{\eta}) = 0$. □

4. APPROXIMATING PROBLEMS

The construction of the approximating problems, required by the theory of consistent approximations described in Section 2, involves the construction of a family of finite-dimensional subspaces, approximating cost functions, and approximating constraint sets. Our selection is largely determined by the fact that we propose to use explicit fixed stepsize Runge-Kutta (RK) [7,18] methods for integrating the dynamic equations (3.1a).

Finite Dimensional Initial-State-Control Subspaces. In Section 3, our optimal control problem **CP** was defined on the normed space $H_{\infty,2}$. Given $N \geq 1$, we will define the corresponding approximating problems, $\mathbf{CP}_N$ on finite dimensional subspaces $H_N = \mathbb{R}^{n_x} \times L_N \subset H_{\infty,2}$, where the $L_N \subset L^m_{\infty,2}[0,1]$ are finite-dimensional spaces spanned by piecewise-continuous functions.

Given an explicit, fixed stepsize RK integration method, we will impose, at the outset, two constraints on the selection of the subspaces L_N:

(i) For any a bounded subset S of $\mathbf{B}$, the RK integration method must give at least first order accuracy, uniformly, in solving the differential equation (3.1), for any $\eta \in S \cap H_N$.

(ii) The data used by the RK integration method is an initial state and a set of control samples[†]. We will require that each set of control samples corresponds to a unique element $u \in L_N$.

The first constraint will be needed to prove that our approximating problems epiconverge to the original problems. For the subspaces L_N we present, we will actually be able to prove more than first order accuracy. The second constraint is imposed to facilitate the definition of the approximating problems and make it possible to define gradients for the approximating cost and constraint functions.

We will now show how explicit, fixed stepsize RK integration methods affect the selection of the subspaces L_N. The generic explicit, fixed stepsize, s-stage RK method computes an approximate solution to a differential equation of the form

$$\dot{x}(t) = \tilde{h}(t, x(t)), \quad x(0) = \xi, \quad t \in [0,1], \tag{4.1a}$$

where $\tilde{h} : \mathbb{R} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ is continuous. It does so by solving the difference equation

$$\bar{x}_{k+1} = \bar{x}_k + \Delta \sum_{i=1}^{s} b_i K_{k,i}, \quad \bar{x}_0 = x(0) = \xi, \quad k \in \mathcal{N} \triangleq \{0, 1, \ldots, N-1\}, \tag{4.1b}$$

with $\Delta = 1/N$, $t_k \triangleq k\Delta$, and $K_{k,i}$ defined by the recursion

[†] The term control samples will be clarified shortly.

$$K_{k,i} = \tilde{h}\,(t_k + c_i\Delta, \bar{x}_k + \Delta\sum_{j=1}^{i-1} a_{i,j}K_{k,j})\,, \quad i \in \mathbf{s}\,, \tag{4.1c}$$

where, according to our notation, $\mathbf{s} \triangleq \{\, 1, \ldots, s \,\}$. The variable $\bar{x}_k$ is the computed estimate of $x\,(t_k)$.

The parameters $a_{i,j}$, c_i and b_i, in (4.1b) and (4.1c) determine the RK method. These parameters are collected in the Butcher array $\mathbf{A} = [c\,, A\,, b\,]$. The Butcher array is often displayed in the form:

$$\mathbf{A} \;=\; \begin{array}{c|cccc} c_1 & 0 & & & \\ c_2 & a_{2,1} & 0 & & \\ \vdots & & \ddots & \ddots & \\ c_s & a_{s,1} & & a_{s,s-1} & 0 \\ \hline & b_1 & \cdots & b_{s-1} & b_s \end{array}$$

The following assumption on the Butcher array parameters will be assumed to hold throughout this paper:

Assumption 4.1. *(a)* For all $i \in \mathbf{s}$, $c_i \in [0,1]$, *(b)* for all $i \in \mathbf{s}$, $b_i > 0$ and $\sum_{i=1}^{s} b_i = 1$. □

Remark 4.2. The condition $\sum_{i=1}^{s} b_i = 1$ is satisfied by all convergent RK methods. Other conditions must be satisfied to achieve higher order convergence for multi-stage RK methods. The condition $b_i > 0$ will be weakened slightly in the sequel.

Now, in our case, $\tilde{h}\,(t\,,x) = h\,(x\,,u\,(t\,))$, and the elements $u\,(t\,)$ of the subspaces L_N will be allowed to be discontinuous from the left at the points $t = t_k + c_i\Delta$. To obtain an accurate integration method for such functions, the values $u\,(t_k + c_i\Delta)$ must sometimes be replaced by left limits, as appropriate for the particular choice of the subspace L_N. We will refer to these values as "control samples" and denote them by $u\,[\tau_{k,i}]$, where, for convenience, we have defined $\tau_{k,i} \triangleq t_k + c_i\Delta$. Specifically, for $\tau \in [0,1]$,

$$u\,[\tau] \triangleq \lim_{t \uparrow \tau} u\,(t\,)\,. \tag{4.2}$$

Clearly if $u\,(\cdot)$ is continuous at $t_k + c_i\Delta$, then $u\,[t_k + c_i\Delta] = u\,(t_k + c_i\Delta)$. Equation (4.1c) evaluates $\tilde{h}\,(\cdot,\cdot)$ s times for each timestep $k \in \mathcal{N}$. So, if we collect those s samples into the matrix $\omega_k \triangleq (u\,[\tau_{k,1}] \cdots u\,[\tau_{k,s}])$, we can replace equations (4.1b) and (4.1c) with

$$\bar{x}_{k+1} = \bar{x}_k + \Delta\sum_{i=1}^{s} b_i K_{k,i}\,, \quad \bar{x}_0 = x\,(0) = \xi\,, \quad k \in \mathcal{N}, \tag{4.3a}$$

where $K_{k,i} \triangleq K_i\,(\bar{x}_k\,, \omega_k)$ which is defined by

$$K_i\,(x\,,\omega) = h\,(x + \Delta\sum_{j=1}^{i-1} a_{i,j}K_j\,(x\,,\omega)\,, \omega^i)\,, \quad i \in \mathbf{s}\,. \tag{4.3b}$$

where ω^i is the i-th column of ω.

We will define the space of control samples, L_N, in such a way that there is a one-to-one correspondence between elements of $u \in L_N$ and the samples of u used by the RK method. The definition of L_N is somewhat complicated by the fact that some of the c_i elements of the Butcher array may have the same value. This causes the RK method to use samples at times $t_k + c_i \Delta$ more than once: a fact that results in a reduction of the dimension of the required space of control samples.

To keep track of the distinct values of the c elements of the Butcher array, we define the ordered set of indices

$$I \triangleq \{ i_1, i_2, \ldots, i_r \} \triangleq \{ i \in \mathbf{s} \mid c_j \neq c_i ,\ \forall j \in \mathbf{s},\ j < i \} , \tag{4.4a}$$

and let

$$I_j \triangleq \{ i \in \mathbf{s} \mid c_i = c_{i_j},\ i_j \in I \} , \quad j \in \mathbf{r} . \tag{4.4b}$$

Thus, the total number of distinct values taken by the elements c_i in the Butcher array is r. For example, if $c = \{ 0, 1/2, 1/2, 1 \}$ then $r = 3$, $I = \{ i_1 = 1, i_2 = 2, i_3 = 4 \}$, $I_1 = \{ 1 \}$, $I_2 = \{ 2, 3 \}$, and $I_3 = \{ 4 \}$. If each c_i is distinct then $r = s$. Otherwise, $r < s$.

Clearly, the r distinct sampling times in the interval $[t_k, t_{k+1}]$, $k \in \mathcal{N}$ are given by τ_{k, i_j}, $j \in \mathbf{r}$, $i_j \in I$. Corresponding to each sampling time there is a control sample $u[\tau_{k, i_j}]$. The collection of these control samples is an element, denoted by $\bar{u}$, of the space $\mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$. We will partition the vectors $\bar{u} \in \mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$ into N blocks, $\bar{u}_k$, consisting of r vectors $\bar{u}_k^j$, of dimension m, *i.e.*,

$$\bar{u} = (\bar{u}_0, \bar{u}_1, \ldots, \bar{u}_{N-1}) , \tag{4.5a}$$

where each $\bar{u}_k$, $k \in \mathcal{N}$ in turn, is of the form

$$\bar{u}_k = (\bar{u}_k^1, \ldots, \bar{u}_k^r) . \tag{4.5b}$$

When convenient for the performing linear algebra, we will consider the elements $\bar{u} \in \mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$, as $Nr \times m$ matrices, *i.e.*, we will identify $\mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$ with the space, $\mathbb{R}^{Nr \times m}$, of $Nr \times m$ matrices. Similarly, we will identify $\mathbb{R}^r \times \mathbb{R}^m$ with $\mathbb{R}^{r \times m}$.

Let G be the $r \times s$ matrix defined by

$$G = \begin{bmatrix} \mathbf{1}_1^T & & & \\ & \mathbf{1}_2^T & & \\ & & \ddots & \\ & & & \mathbf{1}_r^T \end{bmatrix} \tag{4.5c}$$

where, for $j \in \mathbf{r}$, $\mathbf{1}_j^T$, is a row vector of $| I_j |$ ones ($| I_j |$ is the number of elements of I_j). Then the relationship between the components $\bar{u}_k$, $k \in \mathcal{N}$ of a vector $\bar{u} \in \mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$, with $\bar{u}_k^j = u[\tau_{k, i_j}] \in \mathbb{R}^m$, for $j \in \mathbf{r}$ and $i_j \in I$, and the vectors ω_k used by the RK method (4.2a,b) is given by $\omega_k = \bar{u}_k G$.

We are now ready to present two control representations that define subspaces $L_N = L_N^k$, $k = 1,2$, $N \in \mathbb{N}$, such that $\cup_{N=1}^{\infty} L_N$ is dense in $L_2^m[0,1]$. Both representations reduce to simple square pulses for Euler's method (r=1). In addition, we will define two finite dimensional spaces of the form

$$\bar{L}_N^k \triangleq \{ \mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m, \langle \cdot,\cdot \rangle_{\bar{L}_N^k},\ \|\cdot\|_{\bar{L}_N^k} \},\quad k = 1,2,\ N \in \mathbb{N}, \tag{4.5d}$$

consisting of the elements $\bar{u}$ of $\mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$, but with inner products and norms chosen so as to coincide with the corresponding operations in the spaces L_N^k, $k = 1,2$. The spaces $\bar{L}_N^k$ will be needed in defining gradients for the cost and constraint functions of the approximating problems as well as in setting up numerical implementations of optimal control algorithms.

Representation (R1): Piecewise r-th order polynomials.

Let the pulse functions be defined by

$$\Pi_{N,k}^1(t) = \begin{cases} 1 & \text{if } t \in [t_k, t_{k+1}),\ k = 0,\dots,N-2, \\ 1 & \text{if } t \in [t_k, t_{k+1}],\ k = N-1, \\ 0 & \text{elsewhere}, \end{cases} \tag{4.6a}$$

and let

$$L_N^1 \triangleq \{ u \in L_2^m[0,1] \mid u(t) = \sum_{k=0}^{N-1} u_k(t) \Pi_{N,k}^1(t),\ \forall t \in [0,1] \}, \tag{4.6b}$$

where $u_k(t)$ is the vector polynomial

$$u_k(t) \triangleq \sum_{j=0}^{r-1} \beta_{k,j} \left[\frac{t - t_k}{\Delta} \right]^j = \beta_k P(t - t_k), \tag{4.6c}$$

where $\beta_{k,j} \in \mathbb{R}^m$, $\beta_k \triangleq (\beta_{k,0} \cdots \beta_{k,r-1})$ is the $m \times r$ matrix with columns $\beta_{k,j}$, and

$$P(t) \triangleq [1\ \ t/\Delta \cdots (t/\Delta)^{r-1}]^T. \tag{4.6d}$$

Proposition 4.3. Let L_N^1 be defined as in (4.6b) and let $V_{\mathbf{A},N} : L_N^1 \to \mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$ be defined by $V_{\mathbf{A},N}(u) = \bar{u}$, with $\bar{u}_k^j = u[\tau_{k,i_j}]$, $i_j \in I$, $j \in \mathbf{r}$, $k \in \mathcal{N}$. Suppose Assumption 4.1*(a)* holds. Then $V_{\mathbf{A},N}$ is a linear, invertible map.

Proof. From (4.2), (4.6a,b,c) and Assumption 4.1*(a)* if follows that

$$\bar{u}_k^j = u[t_k + c_{i_j}\Delta] = u_k(t_k + c_{i_j}\Delta) = \beta_k P(c_{i_j}\Delta),\quad k \in \mathcal{N}, j \in \mathbf{r}, i_j \in I, \tag{4.7}$$

even if $c_{i_j} = 1$ because $u[t_k + \Delta] = \lim_{t \uparrow t_k + \Delta} u(t) = u_k(t_k + \Delta)$. Thus, for each $k \in \mathcal{N}$, we can rewrite (4.7) in matrix form, as $\bar{u}_k = \beta_k T^{-1}$ (we are thinking of $\bar{u}_k$ as a matrix), where

$$T^{-1} = \left[P(c_{i_1}\Delta)\ P(c_{i_2}\Delta) \cdots P(c_{i_r}\Delta) \right] = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ c_{i_1} & c_{i_2} & & c_{i_r} \\ \vdots & & \ddots & \\ c_{i_1}^{r-1} & c_{i_2}^{r-1} & & c_{i_r}^{r-1} \end{bmatrix} . \tag{4.8}$$

The matrix T^{-1} is a Vandermonde matrix and the r values c_{i_j}, $i_j \in I$, are distinct. Therefore, T^{-1} is non-singular. Hence, for each $k \in \mathcal{N}$, $u_k(t) = \bar{u}_k\, T\, P(t - t_k)$. It follows that $V_{\mathbf{A},N}$ is linear and invertible. □

To complete the definition, in (4.5d), of the spaces $\bar{L}_N^1$ we will now define the required inner product and norm. We define the inner product between two vectors $\bar{u}, \bar{v} \in \bar{L}_N^1$, with $u = V_{\mathbf{A},N}^{-1}(\bar{u})$ and $v = V_{\mathbf{A},N}^{-1}(\bar{v})$, by

$$\begin{aligned} \langle \bar{u}, \bar{v} \rangle_{\bar{L}_N^1} = \langle u, v \rangle_2 &= \sum_{k=0}^{N-1} \int_0^{\Delta} \langle u(t_k + t), v(t_k + t) \rangle\, dt \\ &= \sum_{k=0}^{N-1} \int_0^{\Delta} \langle \bar{u}_k\, T\, P(t), \bar{v}_k\, T\, P(t) \rangle\, dt \\ &= \Delta \sum_{k=0}^{N-1} \operatorname{trace}\!\left(\bar{u}_k\, T\, \frac{1}{\Delta} \int_0^{\Delta} P(t) P(t)^T dt\, T^T \bar{v}_k^T \right), \\ &= \Delta \sum_{k=0}^{N-1} \operatorname{trace}\!\left(\bar{u}_k\, M_1\, \bar{v}_k^T \right), \end{aligned} \tag{4.9a}$$

where T was defined by (4.8), $P(\cdot)$ was defined in (4.6d) and

$$M_1 \triangleq T\, \frac{1}{\Delta} \int_0^{\Delta} P(t) P(t)^T dt\, T^T = T\, \mathrm{Hilb}(r)\, T^T , \tag{4.9b}$$

where

$$\mathrm{Hilb}(r) = \begin{bmatrix} 1 & 1/2 & 1/3 & & 1/r \\ 1/2 & 1/3 & 1/4 & \cdots & 1/r{+}1 \\ 1/3 & 1/4 & 1/5 & & \\ & \vdots & & \ddots & \\ 1/r & 1/r{+}1 & & & 1/2r{+}1 \end{bmatrix}_{r \times r} \tag{4.9c}$$

is the Hilbert matrix, whose i,j-th entry is $1/i{+}j{-}1$. Note that both $\mathrm{Hilb}(r)$ and T are ill-conditioned matrices. However, the product in (4.9c) is well-conditioned (the product corresponds to switching from a power-series polynomial representation of the piecewise polynomials to a Lagrange expansion). The matrix M_1 is symmetric, positive definite because $\mathrm{Hilb}(r)$ is positive definite and T is non-singular.

Remark 4.4. A special class of functions within representation **R1** is the subspace of r-th order, m dimensional splines. The dimension of the spline subspace is only a fraction of the dimension of L_N^1. Our results for **R1** are extended to splines in Appendix A2.

Representation (R2): Stepwise constant functions.

For $j \in \mathbf{r}$, I_j defined in (4.4b), let

$$\tilde{b}_j \triangleq \sum_{i \in I_j} b_i , \tag{4.10a}$$

$$d_j \triangleq \Delta \sum_{i=1}^{j} \tilde{b}_i \ , \quad d_0 \triangleq 0 . \tag{4.10b}$$

If all the c_i elements of the Butcher array have distinct values then $d_j = \Delta \sum_{i=1}^{j} b_i$. At this point, we can replace Assumption 4.1*(b)* with the following weaker assumption:

Assumption 4.1 *(b')* For all $j \in \mathbf{r}$, $\tilde{b}_j > 0$ and $d_r = \Delta$. □

Note that Assumption 4.1*(b')* implies that for all $j \in \mathbf{r}$, $d_j > d_{j-1}$, and that $t_k + d_j \in [t_k, t_{k+1}]$, $k \in \mathcal{N}$.

We introduce an additional assumption which does not rule out any standard RK methods.

Assumption 4.5. For $j \in \mathbf{r}$ and $i_j \in I$, $d_{j-1} \le c_{i_j}\Delta \le d_j$, so that $\tau_{k,i_j} \in [t_k + d_{j-1}, t_k + d_j]$. □

We now define the pulse functions

$$\Pi^2_{N,k,j}(t) = \begin{cases} 1 & \text{if } t \in [t_k + d_{j-1}, t_k + d_j), \ \ k \in \mathcal{N}, \ \ j \in \mathbf{r}, \\ 0 & \text{elsewhere} \ , \end{cases} \tag{4.11a}$$

with $\Pi^2_{N,N,r}$ closed. With $\beta_{k,j} \in \mathbb{R}^m$, let

$$L^2_N \triangleq \{ u \in L^m_2[0,1] \mid u(t) = \sum_{k=0}^{N-1} \sum_{j=1}^{r} \beta_{k,j} \, \Pi^2_{N,k,j}(t), \ \ \forall t \in [0,1] \} . \tag{4.11b}$$

Proposition 4.6. Let L^2_N be defined as in (4.11b) and let $V_{\mathbf{A},N} : L^2_N \to \mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$ be defined by $V_{\mathbf{A},N}(u) = \bar{u}$, with $\bar{u}^j_k = u[\tau_{k,i_j}]$, $j \in \mathbf{r}$, $i_j \in I$, $k \in \mathcal{N}$. Suppose Assumptions 4.1*(b')* and 4.5 hold. Then $V_{\mathbf{A},N}$ is a linear, invertible map.

Proof. Assumption 4.1*(b')* implies that any $u \in L^2_N$ is specified by a unique set of coefficients $\beta_{k,j}$. Assumption 4.5 implies that $\bar{u}^j_k = \beta_{k,j}$ even if $c_{i_j}\Delta = d_i$ since $u[\tau_{k,i_j}]$ is defined as the left limit of $u(\tau_{k,i_j})$. Thus, $V_{\mathbf{A},N}$ is invertible; linearity of $V_{\mathbf{A},N}$ is obvious. □

To complete the definition, in (4.5d), of the spaces $\bar{L}^2_N$ we will now define the required inner product and norm. We define the inner product between two vectors $\bar{u}, \bar{v} \in \bar{L}^2_N$, with $u = V^{-1}_{\mathbf{A},N}(\bar{u})$ and $v = V^{-1}_{\mathbf{A},N}(\bar{v})$, by

$$\langle \bar{u}, \bar{v} \rangle_{\bar{L}_N^2} = \langle u, v \rangle_2 = \sum_{k=0}^{N-1} \sum_{j=1}^{r} \int_{d_{j-1}}^{d_j} \langle u(t_k + t), v(t_k + t) \rangle dt$$

$$= \Delta \sum_{k=0}^{N-1} \sum_{j=1}^{r} \tilde{b}_j \langle \bar{u}_k^j, \bar{v}_k^j \rangle dt$$

$$= \Delta \sum_{k=0}^{N-1} \operatorname{trace}(\bar{u}_k M_2 \bar{v}_k) , \tag{4.12a}$$

where,

$$M_2 = \begin{bmatrix} \tilde{b}_1 & & 0 \\ & \ddots & \\ 0 & & \tilde{b}_r \end{bmatrix} . \tag{4.12b}$$

Since all $\tilde{b}_j > 0$, M_2 is diagonal, positive definite. Given $\bar{u} \in \bar{L}_N^2$, its norm is $\|\bar{u}\|_{\bar{L}_N^2}^2 = \langle \bar{u}, \bar{u} \rangle_{\bar{L}_N^2}$.

Remark 4.7. In place of (4.10b), we could have used the alternate definition $d_j \triangleq \Delta \sum_{i=1}^{j} b_i$ and set $\bar{u}_k^j = u[\tau_{k,j}]$ for all $j \in \mathbf{s}$, $k \in \mathcal{N}$. In this way, samples corresponding to repeated values of c_j in the Butcher array would be treated as independent values and the space L_N would have to be correspondingly enlarged. However, Proposition 6.1 in Section 6 indicates that (4.10b) is the preferable definition.

Definition of Approximating Problems. For $N \in \mathbb{N}$, let

$$H_N \triangleq \mathbb{R}^{n_s} \times L_N , \tag{4.13a}$$

where $L_N = L_N^1$ for representation **R1** or $L_N = L_N^2$ for representation **R2**. $H_N \subset H_{\infty,2}$ inherits the inner product from $H_{\infty,2}$ which, for $\eta', \eta'' \in H_N$, with $\eta' = (\xi', u')$ and $\eta'' = (\xi'', u'')$, is given by

$$\langle \eta', \eta'' \rangle_H \triangleq \langle \xi', \xi'' \rangle + \langle u', u'' \rangle_2, \tag{4.13b}$$

and hence for any $\eta \in H_N$, $\|\eta\|_H^2 = \langle \eta, \eta \rangle_H$. Similarly, we define the spaces $\bar{H}_N$ by

$$\bar{H}_N \triangleq \mathbb{R}^{n_s} \times \bar{L}_N , \tag{4.14a}$$

where $\bar{L}_N = \bar{L}_N^1$ or $\bar{L}_N = \bar{L}_N^2$. The inner product on $\bar{H}_N$ is defined by

$$\langle \bar{\eta}', \bar{\eta}'' \rangle_{\bar{H}_N} \triangleq \langle \xi', \xi'' \rangle + \langle \bar{u}', \bar{u}'' \rangle_{\bar{L}_N}, \tag{4.14b}$$

and the norm correspondingly. Let $W_{\mathbf{A},N} : H_N \to \bar{H}_N$ be defined by $W_{\mathbf{A},N}(\eta) = (\xi, V_{\mathbf{A},N}(u))$, where $\eta = (\xi, u)$. Then we see that $W_{\mathbf{A},N}$ is a linear, nonsingular map, and, with our definition of the norms on $\bar{H}_N$, provides an isometric isomorphism between H_N and $\bar{H}_N$. Thus, we can use the spaces H_N and $\bar{H}_N$ interchangeably.

We now define control constraint sets for the approximating problems. Let U be the convex, compact set used to define $\mathbf{U}$ in (3.3a). Then, we define

$$\overline{\mathbf{U}}_N \triangleq \{ \bar{u} \in \bar{L}_N \mid \bar{u}_k^j \in U \; \forall \, k \in \mathcal{N}, j \in \mathbf{r} \} \tag{4.15a}$$

$$\overline{\mathbf{H}}_N \triangleq \mathbb{R}^{n_x} \times \overline{\mathbf{U}}_N \subset \bar{H}_N \, , \tag{4.15b}$$

$$\mathbf{H}_N \triangleq \mathbb{R}^{n_x} \times V_{\mathbf{A},N}^{-1}(\overline{\mathbf{U}}_N) \subset H_N \, . \tag{4.15c}$$

We assume that $\rho_{\max}$ was chosen large enough in (3.3c) so that $\mathbf{H}_N \subset \mathbf{B}$.

Next, with $\eta \in H_N$ and $\bar{\eta} = W_{\mathbf{A},N}(\eta)$, we will denote the solutions of (4.3a,b) by $\{ \bar{x}_k^{\eta} \}_{k=0}^{N}$ or, equivalently, $\{ \bar{x}_k^{\bar{\eta}} \}_{k=0}^{N}$. The variable $\bar{x}_k^{\eta}$ is thus the computed estimate of $x^{\eta}(t_k)$. Finally, let

$$f_N^{\nu}(\eta) \triangleq \zeta^{\nu}(\xi, \bar{x}_N^{\eta}) \equiv \bar{f}_N^{\nu}(\bar{\eta}) \triangleq \zeta^{\nu}(\xi, \bar{x}_N^{\bar{\eta}}) \; , \quad \nu \in \mathbf{q} \, , \tag{4.16}$$

where $\zeta^{\nu}(\cdot,\cdot)$ was used to define $f^{\nu}(\eta)$ in (3.4c). Then we can state the approximating problems as:

$$\mathbf{CP}_N \qquad \min_{\eta \in \mathbf{H}_N} \; \{ \psi_{o,N}(\eta) \mid \psi_{c,N}(\eta) \le 0 \} \, , \tag{4.17}$$

where $\psi_{o,N}(\eta) \triangleq \max_{\nu \in \mathbf{q}_o} f_N^{\nu}(\eta)$ and $\psi_{c,N}(\eta) \triangleq \max_{\nu \in \mathbf{q}_c + q_o} f_N^{\nu}(\eta)$.

Note that for any $\eta \in \mathbf{H} \cap H_N$, where $\mathbf{H}$ was defined in (3.3b), $W_{\mathbf{A},N}(\eta) \in \overline{\mathbf{H}}_N$ because $u(t) \in U$ for all $t \in [0,1]$ implies that $u[\tau_{k,j}] \in U$ for all $k \in \mathcal{N}$, $j \in \mathbf{s}$. By (4.15c), this implies that $\mathbf{H} \cap H_N \subset \mathbf{H}_N$. Unfortunately, for representation **R1**, if $r \ge 2$ (except for the case $r = 2$ and the Butcher array elements $c = \{0,1\}$), $\mathbf{H}_N \ne \mathbf{H} \cap H_N$ because, given $\bar{u} \in \bar{L}_N^1$, generally $\| V_{\mathbf{A},N}(\bar{u}) \|_{\infty} > \| \bar{u} \|_{\infty}$, [4, p. 25]. Hence, if $\{ \eta_N \}_{N \in \mathbf{N}}$, $\mathbf{N} \subset \mathbb{N}$, is a sequence of approximate solutions to the problems $\mathbf{CP}_N$, it is possible for any η_N, $N \in \mathbf{N}$, to violate the control constraints. However, as we will see, the limit points of such a sequence must satisfy the control constraints. This problem could have been avoided by choosing $\mathbf{H}_N \triangleq \mathbf{H} \cap H_N$ (as in [24]) and letting $\overline{\mathbf{H}}_N \triangleq W_{A,N}(\mathbf{H}_N)$, but the set $\overline{\mathbf{H}}_N$ would be difficult to characterize. For representation **R2** (or **R1**, $r = 1$, or $r = 2$ and $c = \{0,1\}$), $\bar{u} \in \overline{\mathbf{U}}_N \Leftrightarrow V_{\mathbf{A},N}^{-1}(\bar{u}) \in \mathbf{U}$.

Nesting. The theory of consistent approximations is stated in terms of nested subspaces H_N. This allows the approximate solution of an approximating problem $\mathbf{CP}_{N_1}$ to be used as a "warm-start" for an approximating problem $\mathbf{CP}_{N_2}$ with a higher discretization level ($N_2 > N_1$) (see [17,25]).

For representation **R1**, $L_N \subset L_{2N}$ so that doubling the discretization level nests the subspaces. If $u \in L_N$, then $\bar{v} = V_{\mathbf{A},2N}(u)$ can be determined from $\bar{u} = V_{\mathbf{A},N}(u)$ using (4.6c) and (4.8): for $k \in \mathcal{N}$ and $j \in \mathbf{r}$, $\bar{v}_{2k}^j = \bar{u}_k \, T \, P(c_j / 2N)$ and $\bar{v}_{2k+1}^j = \bar{u}_k \, T \, P((c_j + 1)/2N)$. For representation **R2**, $L_N \subset L_{dN}$ where d is the smallest common denominator for all of the b_j, $j \in \mathbf{s}$ in the Butcher array, which is finite assuming, as is typically the case, that all b_j are rational. Thus, the discretization level must be increased by factors of d to achieve nesting. If $u \in L_N$ and $\bar{u} = V_{\mathbf{A},N}(u)$, then $\bar{v} = V_{\mathbf{A},2N}(u)$ is given, for $k \in \mathcal{N}$, $i,j \in \mathbf{r}$, and $l = 1, \ldots, d$, by $\bar{v}_{dk+l}^i = \bar{u}_k^j$ for $d_{j-1} \le l/d < d_j$, where d_j is defined in (4.10b).

Epiconvergence. We are now ready to establish the epiconvergence of the approximating problems. First we present convergence and continuity properties for the solutions computed by Runge-Kutta integration on $\overline{H}_N$. The proof of the following lemma is given in Appendix A1.

Lemma 4.8. For representation **R1**, suppose that Assumptions 3.1*(a)*, 4.1 hold. For representation **R2**, suppose that Assumptions 3.1*(a)*, 4.1, and 4.5 hold.

(i) *Convergence.* For any bounded subset $S \subset \mathbf{B}$, there exists $\kappa < \infty$ and $N^* < \infty$, such that for any $\eta \in S \cap H_N$ and $N \geq N^*$,

$$\| x^{\eta}(t_k) - \bar{x}_k^{\eta} \| \leq \frac{\kappa}{N} \ , \quad k \in \mathcal{N} \cup \{ N \} . \tag{4.18a}$$

Additionally, if the Runge-Kutta method is order ρ, (see [7,18]) and $h(\cdot,\cdot)$ is $\rho-1$ times Lipschitz continuously differentiable, then for representation **R1**, there exists $\kappa < \infty$ and $N^* < \infty$, such that for any $\eta \in S \cap H_N$ and $N \geq N^*$,

$$\| x^{\eta}(t_k) - \bar{x}_k^{\eta} \| \leq \frac{\kappa}{N^{\rho}} \ , \quad k \in \mathcal{N} \cup \{ N \} . \tag{4.18b}$$

The same result hold for representation **R2** if $h(x,u) = \tilde{h}(x) + Bu$ where B is an $n \times m$, constant matrix.
(ii) *Lipschitz Continuity.* The solutions $\{ \bar{x}_k^{\eta} \}_{k=0}^{N}$ are Lipschitz continuous in η on bounded sets, uniformly in k. That is, for any $\eta, \eta' \in S \cap H_N$ with $S \subset \mathbf{B}$ bounded, there exists $\kappa < \infty$, independent of η, such that for $\bar{\eta} = W_{\mathbf{A},N}(\eta)$ and $\bar{\eta}' = W_{\mathbf{A},N}(\eta')$

$$\max_{k \in \mathcal{N} \cup \{ N \}} \| \bar{x}_k^{\eta} - \bar{x}_k^{\eta'} \| \leq \kappa \| \bar{\eta} - \bar{\eta}' \|_{\overline{H}_N} . \tag{4.18c}$$

□

In proving consistency, we will need to add a version of Slater's constraint qualification on the problem **CP**.

Assumption 4.9. For every $\eta \in \mathbf{H}$ such that $\psi_c(\eta) \leq 0$, there exists a sequence $\{ \eta_N \}_{N=1}^{\infty}$ such that $\eta_N \in \mathbf{H}$, $\psi_c(\eta_N) < 0$, and $\eta_N \to \eta$ as $N \to \infty$. □

Theorem 4.10. (Epiconvergence) Suppose that Assumptions 3.1*(a)*, 4.1 and 4.9 (and also 4.5 for representation **R2**) hold. Let $\mathbf{N} = \{ d^n \}_{n=1}^{\infty}$ where $d = 2$ for representation **R1** and d is the least common denominator of $\tilde{b}_j$, $j \in \mathbf{s}$, for representation **R2**. Then, the problems $\{ \mathbf{CP}_N \}_{N \in \mathbf{N}}$ converge epigraphically to the problem **CP** as $N \to \infty$.

Proof. Let $S \subset \mathbf{B}$ be bounded. Then, by Assumption 3.1*(b)* and Lemma 4.8*(i)*, there exists $\kappa', \kappa < \infty$ such that for any $\nu \in \mathbf{q}$ and for any $\eta_N \in S \cap \mathbf{H}_N$,

$$| f^{\nu}(\eta_N) - f_N^{\nu}(\eta_N) | = | \zeta^{\nu}(\xi_N, x^{\eta_N}(1)) - \zeta^{\nu}(\xi_N, \bar{x}_N^{\eta_N}) | \leq \kappa' \| x^{\eta_N}(1) - \bar{x}_N^{\eta_N} \| \leq \frac{\kappa}{N} . \tag{4.19a}$$

Now, let $v' \in \mathbf{q}_o$ be such that $\psi_o(\eta_N) = f^{v'}(\eta_N)$. Then,

$$\psi_o(\eta_N) - \psi_{o,N}(\eta_N) = f^{v'}(\eta_N) - \psi_{o,N}(\eta_N) \le f^{v'}(\eta_N) - f_N^{v'}(\eta_N) \le \frac{\kappa}{N} . \tag{4.19b}$$

By reversing the roles of $\psi_o(\eta)$ and $\psi_{o,N}(\eta_N)$ we can conclude that

$$| \psi_o(\eta_N) - \psi_{o,N}(\eta_N) | \le \frac{\kappa}{N} . \tag{4.20a}$$

Similarly,

$$| \psi_c(\eta_N) - \psi_{c,N}(\eta_N) | \le \frac{\kappa}{N} . \tag{4.20b}$$

Now, given $\eta \in \mathbf{H}$ such that $\psi_c(\eta) \le 0$, there exists, by Assumption 4.9, a sequence $S = \{ \eta_N \}_{N=1}^{\infty}$, with $\eta_N \in \mathbf{H}$, such that $\eta_N \to \eta$ as $N \to \infty$ (hence S is bounded), and $\psi_c(\eta_N) < 0$ for all N. Now, clearly for each $N = d^n$, there exists $j_n \in \mathbf{N}$, finite, and $\eta'_{j_n} \in \mathbf{H}_{j_n}$ such that *(a)* $\kappa / j_n \le -½\psi_c(\eta_N)$, *(b)* $\| \eta_{j_n}' - \eta_N \| \le 1/N$, since, for both representations **R1** and **R2**, the union of the subspaces H_N is dense in H_2 which contains $H_{\infty,2}$ and $\mathbf{H} \cap H_N \subset \mathbf{H}_N$, *(c)* $\psi_c(\eta_{j_n}') \le 1/2\psi_c(\eta_N)$ due to Theorem 3.1*(iii)*, and *(d)* $j_n < j_{n+1}$. It follows from (4.20b) that $\psi_{c,j_{n+t}}(\eta_{j_n}') \le \psi_c(\eta_{j_n}') + k / j_n \le ½\psi_c(\eta_N) + k / j_n \le 0$ for any $n, k \in \mathbb{N}$. Now consider the sequence $S'' = \{ \eta_k'' \}_{k=j_1}^{\infty}$ defined as follows: if $k = j_n$ for some n, then $\eta_k'' = \eta_{j_n}'$ for $k = j_n, j_n+1, j_n+2, \ldots, j_{n+1}-1$. Then we see that $\psi_{c,k}(\eta_k'') \le 0$ for all k, $\eta_k'' \to \eta$ as $k \to \infty$ (hence S'' is bounded), and by (4.20a) and Theorem 3.2*(iii)* that $\lim \psi_{o,N}(\eta_N) = \psi_o(\eta)$. Thus, part *(a)* of Definition 2.1 is satisfied.

Now let $S = \{ \eta_N \}_{N \in K}$, $K \subset \mathbf{N}$, be a sequence with $\eta_N = (\xi_N, u_N) \in \mathbf{H}_N$ and $\psi_{c,N}(\eta_N) \le 0$ for all $N \in K$, and suppose that $\eta_N \to^K \eta = (\xi, u)$. For any $v \in \mathbb{R}^m$, let $d(v, U) \triangleq \min_{v' \in U} \| v - v' \|$. For each N, $V_{\mathrm{A},N}(u_N) \in \overline{\mathbf{U}}_N$ so that $\bar{u}_k^j \in U$ for all $k \in \mathcal{N}$, $j \in \mathbf{r}$. Thus, for representation **R1**, $\overline{\lim}_{t \in [0,1], N \in K} d(u_N(t), U) = 0$ since u_N is composed of polynomials with bounded coefficients (hence bounded derivative) defined over progressively smaller intervals. For representation **R2**, $d(u_N(t), U) = 0$ for all $N \in K$ and $t \in [0,1]$ since u_N is piecewise constant. This implies that $u \in \mathbf{U}$; hence $\eta \in \mathbf{H}$. Furthermore, $\psi_c(\eta) \le 0$ by (4.20b) and the continuity of $\psi_c(\cdot)$, and, again by (4.20a), $\lim \psi_{o,N}(\eta_N) = \psi_o(\eta)$. Thus, part *(b)* of Definition holds. □

Remark 4.11. The fact that Theorem 4.10 depends on Assumption 4.1 implies that consistency of the RK method is not enough to ensure epiconvergence of the approximating problems to the original problem. This explains why, as Hager observed in [16], methods, such as the Improved Euler method of integration, with $b_j = 0$ for some j cannot be used for optimal control problem discretization. But also, methods with $b_j < 0$ for some j cannot be used. For example, the third order method with Butcher array

$$\mathbf{A} = \begin{array}{c|ccc} 0 & 0 & & \\ 1/4 & 1/4 & & \\ 1 & -7/5 & 12/5 & \\ \hline & -1/6 & 8/9 & 5/18 \end{array}$$

when used to discretize the problem described in Section 6, results (observed numerically) in cost function approximations that are concave along some directions, while the original cost function is strictly convex. Hence it cannot lead to epiconvergence of the approximating problems to the original problem.

Factors in Selecting the Control Representation. The choice of selecting $L_N = L_N^1$ versus $L_N = L_N^2$ depends on the relative importance of approximation error versus constraint satisfaction. It follows from the proof of epiconvergence, that irrespective of which representation is used, if $\{\eta_N\}_{N \in \mathbb{N}}$ is a sequence such that $\eta_N \in \mathbf{H}_N$, and $\eta_N \to \eta$, then $\eta \in \mathbf{H}$. Thus η satisfies the control constraints. However, as mentioned earlier, if representation **R1** is used, then η_N may not satisfy the control constraints for any finite N. Since a numerical solution must be obtained after a finite number of iterations, except for the case $r = 2$ and $c = \{0, 1\}$, representation **R2** must be used if absolute satisfaction of control constraints is required.

If some violation of control constraints is permissible, then representation **R1** is preferable to representation **R2** because a tighter bound for the error of the approximate solution can be established for **R1** than for **R2**. To see this, let $\eta_N^* = (\xi^*_N, u_N^*)$, $N \in \mathbb{N}$, be a local minimizer of the finite-dimensional problem $\mathbf{CP}_N$. This solution is computed by setting $\eta_N^* = W_{\mathbf{A},N}^{-1}(\bar{\eta}_N^*)$, where $\bar{\eta}_N^*$ is the result of a numerical algorithm implemented on a computer using the formulae to be presented in the next section.

The accuracy of the approximate control solutions u_N^* can be determined as follows. Assume that $u_N^* \to u^*$ as $N \to \infty$ and that u^* is a local minimizer of **CP** (if the u_N^* solutions are uniformly strict minimizers then u^* must be a local minimizer by Theorem 2.2). Let $\bar{u}^* \in \mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$ be such that $\bar{u}^{*j}_k = u^*(\tau_{k,j})$, for $k \in \mathcal{N}, j \in \mathbf{r}$. Then

$$\|u^* - u_N^*\|_2 \le \|u^* - V_{\mathbf{A},N}^{-1}(\bar{u}^*)\|_2 + \|V_{\mathbf{A},N}^{-1}(\bar{u}^*) - u_N^*\|_2 = \|u^* - V_{\mathbf{A},N}^{-1}(\bar{u}^*)\|_2 + \|\bar{u}^* - \bar{u}_N^*\|_{\bar{L}_N} \, . \quad (4.21)$$

The quantity $\|\bar{u}^* - \bar{u}_N^*\|_{\bar{L}_N}$ is not affected by the choice of control representations. For smooth, unconstrained problems discretized by symmetric RK methods, a bound for $\|\bar{u}^* - \bar{u}_N^*\|_{\bar{L}_N}$ can be found in [16, Thm. 3.1] (see Proposition 6.1 in this paper for an improved bound for one particular method). The quantity $\|u^* - V_{\mathbf{A},N}^{-1}(\bar{u}^*)\|_2$ is the error between u^* and the element of L_N^1 or L_N^2 that interpolates $u^*(t)$ at $t = \tau_{k,j}$, $k \in \mathcal{N}$ and $j \in \mathbf{r}$. The piecewise polynomials of representation **R1** are generally better interpolators for $u^*(\cdot)$, expect near non-smooth points, than the functions of **R2**. For $u^*(\cdot)$ sufficiently smooth, $\|u^* - V_{\mathbf{A},N}^{-1}(\bar{u}^*)\|_2$ is of order r for representation **R1** (see [4]), but only first order for representation **R2**.

5. OPTIMALITY FUNCTIONS FOR THE APPROXIMATING PROBLEMS

In order to develop optimality functions for our approximating problems we must determine the gradients of the cost and constraint functions for the approximating problems.

At each time step, the RK integration formula is a function of the current state estimate $\bar{x}_k$ and r control samples $\bar{u}_k = (\bar{u}_k^1, \ldots, \bar{u}_k^r)$. So, let $F : \mathbb{R}^{n_x} \times (\mathbb{R}^r \times \mathbb{R}^m) \to \mathbb{R}^n$ be defined by

$$F(x, w) = x + \Delta \sum_{i=1}^{s} b_i K_i(x, wG)\,, \tag{5.1}$$

where $w = (w^1, \ldots, w^r) \in \mathbb{R}^r \times \mathbb{R}^m$ is being treated as an $m \times r$ matrix, $\omega = wG \in \mathbb{R}^{s \times m}$, with G defined in (4.5c), and $K_i(x, \omega)$ is defined in (4.3b). Then, referring to (4.2a,b), we see that for any $\bar{\eta} = (\xi, \bar{u}) \in \bar{H}_N$, with $\bar{H}_N$ defined in (4.14a)

$$\bar{x}_{k+1}^{\eta} = F(\bar{x}_k, \bar{u}_k)\,, \quad \bar{x}_0 = \xi\,, \quad k \in \mathcal{N}\,. \tag{5.2}$$

The derivative of $F(\cdot, \cdot)$ with respect to the j-th component of w is, with I_j defined in (4.4b), given by

$$\begin{aligned} F_{w^j}(x, w) &= \Delta \frac{\partial}{\partial w^j} \sum_{i=1}^{s} b_i K_i(x, wG) \\ &= \Delta \sum_{l \in I_j} \frac{\partial}{\partial \omega^l} \sum_{i=1}^{s} b_i K_i(x, \omega) \\ &= \Delta \sum_{l \in I_j} \left[b_l h_u(Y_{k,l}, w^l) + \Delta \sum_{i=1}^{s} b_i h_x(Y_{k,i}, w^i) \sum_{p=1}^{i-1} \frac{\partial}{\partial \omega^l} K_p(x, \omega) \right], \end{aligned} \tag{5.3}$$

where $Y_{k,i} \triangleq x + \Delta \sum_{j=1}^{i-1} K_j(x, \omega)$.

Theorem 5.1. Let $N \in \mathbf{N}$, $\eta \in H_N$ and $\bar{\eta} = W_{\mathbf{A},N}(\eta)$. Also, let $\mathbf{M}_N \in \mathbb{R}^{Nr \times Nr}$ be an N-block diagonal matrix defined by

$$\mathbf{M}_N \triangleq \mathrm{diag}[\Delta M, \Delta M, \ldots, \Delta M]\,, \tag{5.4}$$

where $M = M_1$ if $\bar{L}_N = \bar{L}_N^1$ and $M = M_2$ if $\bar{L}_N = \bar{L}_N^2$. Then, for each $\nu \in \mathbf{q}$, the gradient of $f_N^{\nu}(\eta)$ is

$$\nabla f_N^{\nu}(\eta) = \left(\bar{\gamma}_{\xi}^{\nu}(\bar{\eta})\,, V_{\mathbf{A},N}^{-1}(\bar{\gamma}_u^{\nu}(\bar{\eta})\,\mathbf{M}_N^{-1}) \right) \tag{5.5a}$$

where $\bar{\gamma}^{\nu}(\bar{\eta}) = (\bar{\gamma}_{\xi}^{\nu}(\bar{\eta}), \bar{\gamma}_u^{\nu}(\bar{\eta})) \in \bar{H}_N$ is computed according to

$$\bar{\gamma}_{\xi}^{\nu}(\bar{\eta}) = \nabla_{\xi} \zeta^{\nu}(\xi, \bar{x}_N^{\eta}) + \bar{p}_0^{\eta,\nu}\,, \tag{5.5b}$$

$$\bar{\gamma}_u^{\nu}(\bar{\eta})_k^j = F_{w^j}(\bar{x}_k^{\eta}, \bar{u}_k)^T \bar{p}_{k+1}^{\nu,\eta}\,, \quad k \in \mathcal{N}\,, \quad j \in \mathbf{r}\,, \tag{5.5c}$$

with $\bar{p}_k^{\eta,\nu}$ determined by the adjoint equation

$$\bar{p}_k^{\nu} = F_x(\bar{x}_k, \bar{u}_k)^T \bar{p}_{k+1}^{\nu} \ ; \ \bar{p}_N^{\nu} = \zeta_x^{\nu}(\xi, \bar{x}_N)^T \ , \quad k \in \mathcal{N} . \tag{5.5d}$$

The quantities $F_x(\cdot,\cdot)$ and $F_{w^j}(\cdot,\cdot)$ denote the partial derivatives of $F(x, w)$ with respect to x and the j-th component of w.

Proof. First, we note that $V_{\mathbf{A},N}$ as defined on L_N^1 is invertible by Proposition 4.3 and $V_{\mathbf{A},N}$ as defined on L_N^2 is invertible by Proposition 4.6. Next, referring to [23, p. 68], we see that $\bar{\gamma}_{\xi}^{\nu}(\bar{\eta})$ is the gradient of $\bar{f}_N^{\nu}(\bar{\eta})$ with respect to ξ. Similarly, $\bar{\gamma}_u^{\nu}(\bar{\eta})$ is the gradient of $\bar{f}_N^{\nu}(\bar{\eta})$ with respect to $\bar{u} \in \mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$, endowed with the standard l_2 inner product. Hence,

$$Df_N^{\nu}(\eta\,;\delta\eta) = D\bar{f}_N^{\nu}(\bar{\eta}\,;\delta\bar{\eta}) = \langle \bar{\gamma}_{\xi}^{\nu}(\bar{\eta}), \delta\xi \rangle + \langle \bar{\gamma}_u^{\nu}(\bar{\eta}), \delta\bar{u} \rangle_{l_2}$$

$$= \langle \bar{\gamma}_{\xi}^{\nu}(\bar{\eta}), \delta\xi \rangle + \langle \bar{\gamma}_u^{\nu}(\bar{\eta}) \mathbf{M}_N^{-1}, \delta\bar{u} \rangle_{\bar{L}_N} = \langle \gamma_{\xi}^{\nu}(\bar{\eta}), \delta\xi \rangle + \langle V_{\mathbf{A},N}^{-1}(\gamma_u^{\nu}(\bar{\eta}) \mathbf{M}_N^{-1}), \delta u \rangle_{L_N} \ , \tag{5.6}$$

where $\delta\eta = (\delta\xi, \delta u) \in H_N$ and $\delta\bar{\eta} = (\delta\xi, \delta\bar{u}) = W_{\mathbf{A},N}(\delta\eta)$. Since by definition of $\nabla f_N^{\nu}(\eta)$, $Df_N^{\nu}(\eta\,;\delta\eta) = \langle \nabla f_N^{\nu}(\eta), \delta\eta \rangle_H$ for all $\delta\eta \in H_N$, the desired result follows from (5.6). □

Note that for all $k \in \mathcal{N}$, $\gamma_u^{\nu}(\bar{\eta})_k \in \mathbb{R}^r \times \mathbb{R}^m$, and that for all $k \in \mathcal{N}$, $j \in \mathbf{r}$ and $i_j \in I$,

$$\frac{1}{N}\left(\nabla_u f_N^{\nu}(\eta)(\tau_{k,i_1}) \cdots \nabla_u f_N^{\nu}(\eta)(\tau_{k,i_r})\right) M = \left(\bar{\gamma}_u^{\nu}(\bar{\eta})_k^1 \cdots \bar{\gamma}_u^{\nu}(\bar{\eta})_k^r\right) . \tag{5.7}$$

Remark 5.2. At this point, we can draw one very important conclusion. For every $\nu \in \mathbf{q}$, the steepest descent direction, in $\bar{H}_N$, for the function $\bar{f}_N^{\nu}(\cdot)$, at $\bar{\eta}$, is given by $-(\bar{\gamma}_{\xi}^{\nu}(\bar{\eta}), \bar{\gamma}_u^{\nu}(\bar{\eta}) \mathbf{M}_N^{-1})$, and not $-(\bar{\gamma}_{\xi}^{\nu}(\bar{\eta}), \bar{\gamma}_u^{\nu}(\bar{\eta}))$ which is the steepest descent direction that one would obtain with the standard inner product on $\mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$. The naive approach of solving the discrete-time optimal control problem $\mathbf{CP}_N$, using the latter steepest descent directions, amounts to a change of variables that can result in severe ill-conditioning, as we will illustrate in Section 6.

Theorem 5.3. Suppose that Assumptions 3.1, 4.1 (and for representation **R2** 4.5) hold and that the map $V_{\mathbf{A},N}(\cdot)$ is defined as in Proposition 4.3 or Proposition 4.6. Then, the gradients $\nabla f_N^{\nu}(\cdot)$, $\nu \in \mathbf{q}$ are Lipschitz continuous on bounded sets in H_N.

Proof. By Lemma 4.8*(ii)*, the solutions $\bar{x}_k^{\eta}$, of (4.2a,b) are Lipschitz continuous on bounded sets, with respect to $\bar{\eta}$, uniformly in k. A similar argument can be used to show that the adjoint variables $\bar{p}_k^{\eta,\nu}$ are Lipschitz continuous on bounded sets with respect to $\bar{\eta}$, uniformly in k. It follows from Assumption 3.1, (5.5b), and (5.5c) that $\bar{\gamma}^{\nu}(\cdot)$ is Lipschitz continuous on bounded sets. Since $V_{\mathbf{A},N}^{-1}(\cdot)$ is a finite-dimensional, linear operator, it is bounded and hence Lipschitz continuous. Therefore, it is clear from relation (5.5a) that $\nabla f_N^{\nu}(\cdot)$ is Lipschitz continuous on bounded sets. □

We can now define optimality functions for the approximating problems, using the form of the optimality function, presented in (3.9c), for the original problem. For $\mathbf{CP}_N$, we define $\theta_N : H_N \to \mathbb{R}$, with $\sigma > 0$ and the set $\mathbf{H}_N$ is defined in (4.15c), by

$$\theta_N(\eta) \overset{\Delta}{=} \min_{\eta' \in \mathbf{H}_N} \max \left\{ \max_{\nu \in \mathbf{q}_o} \tilde{f}_N^{\nu}(\eta, \eta') - \psi_{o,N}(\eta) - \sigma \psi_{c,N}(\eta)_+ \, , \max_{\nu \in \mathbf{q}_c + q_o} \tilde{f}_N^{\nu}(\eta, \eta') - \psi_{c,N}(\eta)_+ \right\} \qquad (5.8a)$$

where $\psi_{c,N}(\eta)_+ \overset{\Delta}{=} \max \{ 0, \psi_{c,N}(\eta) \}$, and for $\nu \in \mathbf{q}$,

$$\tilde{f}_N^{\nu}(\eta, \eta') \overset{\Delta}{=} f_N^{\nu}(\eta) + \langle \nabla f_N^{\nu}(\eta), \eta' - \eta \rangle_{H_N} + \tfrac{1}{2} \| \eta' - \eta \|_{H_N}^2 \, . \qquad (5.8b)$$

For the purposes of numerical computation [22], we can express (5.8a) in the equivalent form

$$\begin{aligned} \theta_N(\eta) = \min_{\bar{\eta}' \in \bar{\mathbf{H}}_N} \Big\{ & \tfrac{1}{2} \| \bar{\eta}' - \bar{\eta} \|_{\bar{H}_N}^2 \\ & + \max \Big\{ \max_{\nu \in \mathbf{q}_o} \bar{f}_N^{\nu}(\bar{\eta}) + \langle (\bar{\gamma}_\xi^{\nu}(\bar{\eta}), \bar{\gamma}_u^{\nu}(\bar{\eta}) \mathbf{M}_N^{-1}), \bar{\eta}' - \bar{\eta} \rangle_{\bar{H}_N} - \bar{\psi}_{o,N}(\bar{\eta}) - \sigma \bar{\psi}_{c,N}(\bar{\eta})_+ \, , \\ & \max_{\nu \in \mathbf{q}_c + q_o} \bar{f}_N^{\nu}(\bar{\eta}) + \langle (\bar{\gamma}_\xi^{\nu}(\bar{\eta}), \bar{\gamma}_u^{\nu}(\bar{\eta}) \mathbf{M}_N^{-1}), \bar{\eta}' - \bar{\eta} \rangle_{\bar{H}_N} - \bar{\psi}_{c,N}(\bar{\eta})_+ \Big\} \Big\} \end{aligned} \qquad (5.9)$$

where $\bar{\eta} = W_{\mathbf{A},N}(\eta)$ and the set $\bar{\mathbf{H}}_N$ is defined in (4.15b).

It should be obvious that the optimality function is well defined because of the form of the quadratic term and the fact that the minimum is taken over a set of finite dimension. The following theorem confirms that (5.8a) satisfies the definition for an optimality function. The proof is essentially the same as the proof in [3, Thms. 3.6,3.7].

Theorem 5.4. *(i)* For every $\eta \in \mathbf{H}_N$, $\theta_N(\eta) \le 0$; *(ii)* $\theta_N(\cdot)$ is continuous; *(iii)* if $\hat{\eta} \in \mathbf{H}_N$ is a local minimizer for $\mathbf{CP}_N$ then $\theta_N(\hat{\eta}) = 0$. □

Remark 5.5. It can also be shown that $\theta_N(\hat{\eta}) = 0$ if and only if $d_{\eta'} \Psi_N(\hat{\eta}, \hat{\eta} \, ; \eta - \hat{\eta}) \ge 0$ for all $\eta \in \mathbf{H}_N$ where $\Psi_N(\eta, \eta') \overset{\Delta}{=} \max \{ \psi_{o,N}(\eta) - \psi_{o,N}(\eta') - \sigma \psi_{c,N}(\eta')_+ \, , \psi_{c,N}(\eta) - \psi_{c,N}(\eta')_+ \}$. Also, since the matrix $\mathbf{M}_N$ is positive definite for any control representation, it can be seen from (5.9) that if $\hat{\eta}_1$ is a stationary point of $\mathbf{CP}_N$ under representation **R1** and $\hat{\eta}_2$ is a stationary point of $\mathbf{CP}_N$ under representation **R2**, then $W_{\mathbf{A},N}(\hat{\eta}_1) = W_{\mathbf{A},N}(\hat{\eta}_2)$. In particular, the control samples of the stationary points of $\mathbf{CP}_N$ are not affected by the choice of control representations.

Consistency of the Approximations. To complete our demonstration of consistency of approximations we will show that the optimality functions of the approximating problems hypoconverge to the optimality function of the original problem. First we will present a simple algebraic condition which indicates convergence of the gradients. We will need the column vector $\tilde{b} \in \mathbb{R}^r$, with components $\tilde{b}_j$ defined in (4.10a), and the values d_j defined in (4.10b).

Theorem 5.6. For representation **R1**, suppose that Assumptions 3.1 and 4.1 hold. For representation **R2**, suppose that Assumptions 3.1, 4.1, and 4.5 hold. For $N \in \mathbf{N}$, let H_N be defined as in (4.13a), with $L_N = L_N^1$ or $L_N = L_N^2$, and let $f_N^\nu : H_N \to \mathbb{R}$, $\nu \in \mathbf{q}$, be defined by (4.16). Let $M = M_1$ if $L_N = L_N^1$ and let $M = M_2$ if $L_N = L_N^2$. Let S be a bounded subset of $\mathbf{B}$. If

$$M^{-1}\tilde{b} = \mathbf{1}\,, \tag{5.10a}$$

where $\mathbf{1}$ is a vector of r ones, then there exists a $\kappa < \infty$ and an $N^* < \infty$ such that for all $\eta = (\xi, u) \in S \cap H_N$ and $N \geq N^*$,

$$\|\nabla f^\nu(\eta) - \nabla f_N^\nu(\eta)\|_H \leq \frac{\kappa}{N}\,. \tag{5.10b}$$

Proof. To simplify notation, we replace $\bar{x}_k^\eta$ by $\bar{x}_k$, and $\bar{p}_k^{\nu,\eta}$ by $\bar{p}_k^\nu$. Let $S \subset \mathbf{B}$ be bounded and let $\eta = (\xi, u) \in S$. Let $\bar{u} = V_{\mathbf{A},N}(u)$ and $\bar{\eta} = (\xi, \bar{u})$. For each $j \in \mathbf{r}$ and $k \in \mathcal{N}$, $F_{w^j}(\bar{x}_k, \bar{u}_k)$ is given by (5.3). Hence, with $Y_{k,i} \triangleq \bar{x}_k + \Delta \sum_{j=1}^{i-1} a_{i,j} K_j(\bar{x}_k, \omega)$ and $\omega = \bar{u}_k G$, the exists $\kappa_1 < \infty$ such that

$$\begin{aligned}
&\|F_{w^j}(\bar{x}_k, \bar{u}_k) - \Delta \tilde{b}_j h_u(\bar{x}_k, \bar{u}_k^j)\| \\
&\qquad \leq \|F_{w^j}(\bar{x}_k, \bar{u}_k) - \Delta \sum_{l \in I_j} b_l h_u(Y_{k,l}, \bar{u}_k^i)\| + \|\Delta \sum_{l \in I_j} b_l h_u(Y_{k,l}, \bar{u}_k^i) - \Delta \tilde{b}_i h_u(\bar{x}_k, \bar{u}_k^i)\| \\
&\qquad \leq \Delta^2 \| \sum_{l \in I_j} \sum_{i=1}^{s} b_i h_x(Y_{k,i}, \bar{u}_k^i) \sum_{p=1}^{i-1} \frac{\partial}{\partial \omega^l} K_p(\bar{x}_k, \omega)\| + \Delta \sum_{l \in I_j} b_l \|h_u(Y_{k,l}, \bar{u}_k^j) - h_u(\bar{x}_k, \bar{u}_k^j)\| \\
&\qquad \leq \kappa_1 \Delta^2\,,
\end{aligned} \tag{5.11b}$$

where we have used the Lipschitz continuity of $h_u(\cdot, \cdot)$ and the fact that S bounded implies that $\bar{x}_k$ and $\bar{u}_k^j$ are bounded, which implies that for all $j \in \mathbf{r}$, $\|h_u(\bar{x}_k, \bar{u}_k^j)\|$ and $\|h_x(\bar{x}_k, \bar{u}_k^j)\|$ are bounded. Therefore, it follows from (5.5c) that

$$\begin{aligned}
\bar{\gamma}_u^\nu(\bar{\eta})_k &= \left(F_{w^1}(\bar{x}_k, \bar{u}_k)^T \bar{p}_{k+1}^\nu \cdots F_{w^r}(\bar{x}_k, \bar{u}_k)^T \bar{p}_{k+1}^\nu\right) \\
&= \Delta\left(\tilde{b}_1 h_u^T(\bar{x}_k, \bar{u}_k^1)\bar{p}_{k+1}^\nu \cdots \tilde{b}_r h_u^T(\bar{x}_k, \bar{u}_k^r)\bar{p}_{k+1}^\nu\right) + O(\Delta^2)\,.
\end{aligned} \tag{5.11c}$$

Now, from equation (5.5a), $V_{\mathbf{A},N}(\nabla_u f_N(\eta)) = \bar{\gamma}_u^\nu(\bar{\eta})\,\mathbf{M}_N^{-1}$. Therefore, using (5.7) and (5.11c), we obtain

$$V_{\mathbf{A},N}(\nabla_u f_N^\nu(\eta))_k = \frac{\Delta}{\Delta}\left(\tilde{b}_1 h_u(\bar{x}_k, \bar{u}_k^1)^T \bar{p}_{k+1}^\nu \cdots \tilde{b}_r h_u(\bar{x}_k, \bar{u}_k^r)^T \bar{p}_{k+1}^\nu\right)M^{-1} + \frac{O(\Delta^2)}{\Delta}\,. \tag{5.11d}$$

At this point we must deal explicitly with our two control representations. For representation **R1**, $u_N(\cdot)$ is a polynomial on each interval $[t_k, t_{k+1})$. Thus, since S is bounded, for $j, l \in \mathbf{r}$ and $i_j, i_l \in I$, with I defined in (4.4a), there exists $\kappa_2 < \infty$, such that

$$\|\bar{u}_k^j - \bar{u}_k^l\| = \|u[\tau_{k,i_j}] - u[\tau_{k,i_l}]\| \leq \kappa_2 \|\Delta(c_{i_j} - c_{i_l})\| \leq \kappa_2 \Delta \,, \tag{5.12}$$

where Assumption 4.1*(a)* was used to justify the last inequality. Let

$$D \triangleq \left(\tilde{b}_1 h_u^T(\bar{x}_k, \bar{u}_k^1) \bar{p}_{k+1}^{\nu} \cdots \tilde{b}_r h_u^T(\bar{x}_k, \bar{u}_k^r) \bar{p}_{k+1}^{\nu} \right) M^{-1} \,, \tag{5.13a}$$

and let D^j, $j \in \mathbf{r}$, denote the j-th column of D, so that

$$\nabla_u f_N^{\nu}(\eta)[\tau_{k,i_j}] = V_{\mathbf{A},N}(\nabla_u f_N^{\nu}(\eta))_k^j = D^j + O(\Delta) \,, \tag{5.13b}$$

where $\nabla_u f_N^{\nu}(\eta)[\tau_{k,j}]$ is the sample of $\nabla f_N^{\nu}(\eta)(\cdot)$ computed according to (4.2). It follows from Assumptions 3.1*(a)* and 4.1*(b')*, equation (5.12) and the fact that $\bar{p}_{k+1}^{\nu}$ is bounded for any $\eta \in S$, that there exists $\kappa_3, \kappa_4 < \infty$, such that for any $j \in \mathbf{r}$ and $i_j \in I$, and $M_{i,j}^{-1}$ denoting the i,j-th entry of M^{-1},

$$\|D^j - h_u(\bar{x}_k, \bar{u}_k^j)^T p_{k+1}^{\nu} \sum_{i=1}^{r} \tilde{b}_i M_{i,j}^{-1}\| \leq \|\sum_{i=1}^{r} \tilde{b}_i [\, h_u(\bar{x}_k, \bar{u}_k^i) - h_u(\bar{x}_k, \bar{u}_k^j) \,]^T p_{k+1}^{\nu} M_{i,j}^{-1}\|$$

$$\leq \sum_{i=1}^{r} \kappa_3 \|\bar{u}_k^i - \bar{u}_k^j\| \|p_{k+1}^{\nu} M_{i,j}^{-1}\| \leq \kappa_4 \Delta \,, \tag{5.13c}$$

Consequently, if $M^{-1}\tilde{b} = \mathbf{1}$ then $\sum_{i=1}^{r} M_{i,j}^{-1} \tilde{b}_i = 1$ since M is symmetric. Hence for any $j \in \mathbf{r}$,

$$\|D^j - h_u(x_k, u_k^j)^T \bar{p}_{k+1}^{\nu}\| \leq \kappa_4 \Delta \,. \tag{5.13d}$$

Therefore, from (5.13b),

$$\nabla_u f_N^{\nu}(\eta)[\tau_{k,i_j}] = h_u(x_k, \bar{u}_k^j)^T \bar{p}_{k+1}^{\nu} + O(\Delta) \,. \tag{5.13e}$$

For representation **R2**, $\bar{u}_N(\cdot)$ is not Lipschitz continuous on $[t_k, t_{k+1})$, so (5.12) does not hold. However, since $M = M_2$ is diagonal, equation (5.13e) is seen to be true directly from equation (5.11d) if $M^{-1}\tilde{b} = \mathbf{1}$. Now, since S is bounded, *(i)* by Lemmas 4.8*(i)* and A1.3 there exists $\kappa_5 < \infty$ such that $\|\bar{x}_k - x^{\eta}(t_k)\| \leq \kappa_5 \Delta$ and $\|\bar{p}_{k+1}^{\nu} - p^{\eta,\nu}(t_{k+1})\| \leq \kappa_5 \Delta$ and *(ii)* $\bar{p}_{k+1}^{\nu}$ and $h_u(\bar{x}^k, u[\tau_{k,i_j}])$ are bounded. Thus, making use of Theorem 3.2*(v)* and equation (5.13e), the fact that both $x^{\eta}(\cdot)$ and $p^{\eta,\nu}(\cdot)$ are Lipschitz continuous, and $u[\tau_{k,i_j}] = \bar{u}_k^j$, we conclude that there exists $\kappa_6 < \infty$ such that

$$\|\nabla_u f^{\nu}(\eta)[\tau_{k,i_j}] - \nabla_u f_N^{\nu}(\eta)[\tau_{k,i_j}]\| = \|h_u(x^{\eta}(\tau_{k,i_j}), u[\tau_{k,i_j}])^T p^{\eta,\nu}(\tau_{k,i_j}) - h_u(\bar{x}_k, u[\tau_{k,i_j}])^T \bar{p}_{k+1}^{\nu})\|$$

$$\leq \kappa_6 \Delta \,. \tag{5.14b}$$

Now, for $j \in \mathbf{r}$, $i_j \in I$, and $k \in \mathcal{N}$, we have that

$$\|\nabla_u f^{\nu}(\eta)(t) - \nabla_u f_N^{\nu}(\eta)(t)\| \leq \|\nabla_u f^{\nu}(\eta)(t) - \nabla_u f^{\nu}(\eta)[\tau_{k,i_j}]\| + \|\nabla_u f^{\nu}(\eta)[\tau_{k,i_j}] - \nabla_u f_N^{\nu}(\eta)[\tau_{k,i_j}]\|$$

$$+ \|\nabla_u f_N^{\nu}(\eta)[\tau_{k,i_j}] - \nabla_u f_N^{\nu}(\eta)(t)\| \,. \tag{5.15}$$

The second term in (5.15) is order $O(\Delta)$ by (5.14). We will show that the first and third terms in (5.15) are also order $O(\Delta)$. First consider representation **R1**. It follows by inspection of (3.6b) in Theorem 3.2*(v)* that $\nabla_u f^{\nu}(\eta)(\cdot)$ is Lipschitz continuous on $t \in [t_k, t_{k+1})$, $k \in \mathcal{N}$, because $u \in L_N^1$ is Lipschitz continuous on these intervals. Since $\nabla_u f_N^{\nu}(\eta)(\cdot) \in L_N$, it is also Lipschitz continuous on these intervals. Finally, by Assumption 3.1*(a)*, $\tau_{k,i_j} \in [t_k, t_{k+1}]$ for all $k \in \mathcal{N}$. Thus, the first and third terms are of order $O(\Delta)$ for all $t \in [0,1]$. For representation **R2**, $\nabla_u f_N^{\nu}(\eta)(\cdot) \in H_N$ is constant on $t \in [t_k + d_{j-1}, t_k + d_j)$, $j \in \mathbf{r}$ and $k \in \mathcal{N}$. Since $u \in L_N^2$ is constant on these intervals, it again follows by inspection of (3.6b) in Theorem 3.2*(v)* that $\nabla_u f^{\nu}(\eta)(\cdot)$ is Lipschitz continuous on these intervals. Finally, by Assumption 4.5, $\tau_{k,i_j} \in [t_k + d_{j-1}, t_k + d_j]$, for all $k \in \mathcal{N}$ and $j \in \mathbf{r}$. Since $d_0 = 0$ and $d_r = \Delta$, the first and third terms are of order $O(\Delta)$ for all $t \in [0,1]$. We conclude that there exist $\kappa_7 < \infty$ such that

$$\| \nabla_u f^{\nu}(\eta)(t) - \nabla_u f_N^{\nu}(\eta)(t) \| \leq \kappa_7 \Delta\,, \quad t \in [0,1]\,. \tag{5.15b}$$

Next we consider the gradient with respect to initial conditions ξ. From Theorem 3.2*(v)* and (5.5b), $\| \nabla_\xi f^{\nu}(\eta) - \overline{\gamma}_\xi^{\nu}(\eta) \| = \| \nabla_\xi \zeta^{\nu}(\xi, x^{\eta}(1)) - \nabla_\xi \zeta^{\nu}(\xi, \bar{x}_N) \| + \| p^{\nu,\eta}(0) - \bar{p}_0^{\nu} \|$. Thus, since S is bounded, it follows from Assumption 3.1*(b)* and Lemmas 4.8 and A1.3, that there exists $\kappa_8 < \infty$ such that

$$\| \nabla_\xi f^{\nu}(\eta) - \overline{\gamma}_\xi^{\nu}(\eta) \| \leq \kappa_8 \big(\| x^{\eta}(1) - \bar{x}_N \| + \| p^{\nu,\eta}(0) - \bar{p}_0^{\nu} \| \big) \leq \kappa_8 \Delta\,. \tag{5.16}$$

Combining (5.15b) and (5.16), we see that there exists $\kappa < \infty$ such that for any $\eta \in S \cap H_N$,

$$\| \nabla f^{\nu}(\eta) - \nabla f_N^{\nu}(\eta) \|_H \leq \frac{\kappa}{N}\,. \tag{5.17}$$

□

The following proposition states conditions for (5.10a) to hold.

Proposition 5.7.

(a) Suppose $M = M_1$. Then (5.10a) holds if and only if the coefficients of the Butcher array satisfy

$$\sum_{j=1}^{s} b_j c_j^{p-1} = \frac{1}{p}\,, \quad p = 1, \ldots, r\,. \tag{5.18}$$

(b) Suppose $M = M_2$. Then (5.10a) holds if and only if for all $j \in \mathbf{r}$, $\tilde{b}_j > 0$.

Proof. *(a)* For $M = M_1$, it follows from (4.9b) that $M^{-1}\tilde{b} = \mathbf{1}$ if and only if

$$T^{-T}\,\mathrm{Hilb}(s)^{-1} T^{-1} \tilde{b} = \mathbf{1}\,. \tag{5.19a}$$

Now, it can easily be shown that

$$T^{-1}\tilde{b} = \begin{bmatrix} \sum_{j=1}^{r} \tilde{b}_j \\ \sum_{j=1}^{r} \tilde{b}_j c_{i_j} \\ \vdots \\ \sum_{j=1}^{r} \tilde{b}_j c_{i_j}^{r-1} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{s} b_j \\ \sum_{j=1}^{s} b_j \, c_j \\ \vdots \\ \sum_{j=1}^{s} b_j \, c_j^{r-1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \\ \vdots \\ 1/r \end{bmatrix} , \tag{5.19b}$$

where the last equality holds if and only if (5.18) holds. Note that $T^{-1}\tilde{b}$ is then the first column of Hilb(r). Consequently,

$$\text{Hilb}(r)^{-1}T^{-1}\tilde{b} = \text{Hilb}(r)^{-1} \begin{bmatrix} 1 \\ 1/2 \\ \vdots \\ 1/r \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} , \tag{5.19d}$$

which leads us to conclude that

$$M^{-1}\tilde{b} = T^{-T} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & c_{i_1} & \cdots & c_{i_1}^{r-1} \\ 1 & c_{i_2} & & c_{i_2}^{r-1} \\ & & \ddots & \\ 1 & c_{i_r} & \cdots & c_{i_r}^{r-1} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} . \tag{5.19e}$$

(b) Clearly, (5.10a) holds if and only if $M\,\mathbf{1} = \tilde{b}$. For $M = M_2$, if follows from (4.12b) that

$$M\,\mathbf{1} = \begin{bmatrix} \tilde{b}_1 & & \\ & \ddots & \\ & & \tilde{b}_r \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \tilde{b} . \tag{5.20}$$

□

Remark 5.8. The conditions (5.18) on the coefficients of the Butcher array for representation **R1** are necessary conditions for the RK methods to be r-th order accurate [7,18]. The condition with $p = 1$ in (5.18) is the same as the second part of Assumption 4.1*(b')*.

Theorem 5.9. For representation **R1**, suppose that Assumptions 3.1 and 4.1 and equation (5.18) hold and let $d = 2$. For representation **R2**, suppose that Assumptions 3.1, 4.1, and 4.5 hold and let d be the least common denominator for the elements b_j, $j \in \mathbf{s}$ of the Butcher array. Let $\mathbf{N} \triangleq \{ d^n \}_{n=1}^{\infty}$ and suppose that $\{ \eta_N \}_{N \in \mathbf{N}}$ is such that $\eta_N \in \mathbf{H}_N$ for all $N \in \mathbf{N}$ and $\eta_N \to \eta$ as $N \to \infty$. Then $\theta_N(\eta_N) \to \theta(\eta)$ as $N \to \infty$.

Proof. Let $\tilde{\Psi} : \mathbf{H}_c \times \mathbf{H}_c \to \mathbb{R}$ be defined by

$$\tilde{\Psi}(\eta,\eta') \triangleq \max\left\{ \max_{\nu \in q_o} \tilde{f}^{\nu}(\eta,\eta') - \psi_o(\eta) - \sigma\psi_c(\eta)_+ \, , \, \max_{\nu \in q_c + q_o} \tilde{f}^{\nu}(\eta,\eta') - \psi_c(\eta)_+ \right\} , \tag{5.21a}$$

and $\tilde{\Psi}_N : \mathbf{H}_N \times \mathbf{H}_N \to \mathbb{R}$ be defined by

$$\tilde{\Psi}_N(\eta,\eta') \triangleq \max\left\{ \max_{\nu \in q_o} \tilde{f}_N^{\nu}(\eta,\eta') - \psi_{o,N}(\eta) - \sigma\psi_{c,N}(\eta)_+ \, , \max_{\nu \in q_c + q_o} \tilde{f}_N^{\nu}(\eta,\eta') - \psi_{c,N}(\eta) \right\} , \tag{5.21b}$$

so that, $\theta(\eta) = \min_{\eta' \in \mathbf{H}_c} \tilde{\Psi}(\eta,\eta')$, and $\theta_N(\eta) = \min_{\eta' \in \mathbf{H}_N} \tilde{\Psi}_N(\eta,\eta')$. Now, suppose that $\{ \eta_N \}_{N=1}^{\infty}$ is a sequence such that, for all N, $\eta_N \in \mathbf{H}_N$ and $\eta_N \to \eta$. From the proof of Theorem 4.10, $\eta \in \mathbf{H}$. Let $\hat{\eta} \in \mathbf{H}$ be such that $\theta(\eta) = \tilde{\Psi}(\eta,\hat{\eta})$, and let $\{ \eta'_N \}_{N=1}^{\infty}$ be any sequence such that, for all N, $\eta'_N \in \mathbf{H}_N$ and $\eta'_N \to \hat{\eta}$. Then,

$$\theta_N(\eta_N) \le \tilde{\Psi}_N(\eta_N,\eta'_N) \le \tilde{\Psi}(\eta_N,\eta'_N) +$$

$$\max\left\{ \max_{\nu \in q_o} \{ \tilde{f}_N^{\nu}(\eta_N,\eta'_N) - \tilde{f}^{\nu}(\eta_N,\eta'_N) \} - [\psi_{o,N}(\eta_N) - \psi_N(\eta_N)] - [\sigma\psi_{c,N}(\eta_N)_+ - \sigma\psi_c(\eta_N)_+] \, , \right.$$

$$\left. \max_{\nu \in q_c + q_o} \{ \tilde{f}_N^{\nu}(\eta_N,\eta'_N) - \tilde{f}^{\nu}(\eta_N,\eta'_N) \} - [\psi_{c,N}(\eta_N) - \psi_{c,N}(\eta_N)] \right\} \tag{5.22}$$

It follows from Theorem 4.11, Theorem 5.6, Proposition 5.7 and the fact that $\{ \eta_N \}_{N=1}^{\infty}$ is a bounded set, that each part of the max term on the righthand side of (5.22) converges to zero as $N \to \infty$. The quantity $\tilde{\Psi}(\eta_N,\eta'_N)$ converges to $\theta(\eta)$ since $\eta_N \to \eta$, $\eta'_N \to \hat{\eta}$ and $\tilde{\Psi}(\cdot,\cdot)$ is continuous. Thus, taking limits of both sides of equation (5.22), we obtain that $\overline{\lim}\,\theta_N(\eta_N) \le \theta(\eta)$ (this proves that Definition 2.5 holds for the optimality functions of the approximating problems). Now, for all N, let $\hat{\eta}_N \in \mathbf{H}_N$ be such that $\theta_N(\eta_N) = \tilde{\Psi}_N(\eta_N,\hat{\eta}_N)$. Then, $\theta(\eta_N) \le \tilde{\Psi}(\eta_N,\hat{\eta}_N)$ and proceeding in a similar fashion as (5.22) and taking limits, we see that $\theta(\eta) \le \underline{\lim}\,\theta_N(\eta_N)$. Hence, together with the previous result, we can conclude that $\theta_N(\eta_N) \to \theta(\eta)$ as $N \to \infty$. □

The following results is a direct result of Theorem 4.10 (epiconvergence) and Theorem 5.9:

Corollary 5.10. (Consistency) For representation **R1**, suppose that Assumptions 3.1, 4.1 and 4.9 and equation (5.18) hold. For representation **R2**, suppose that Assumptions 3.1, 4.1, 4.5 and 4.9 hold. Let $\mathbf{N} = \{ d^n \}_{n=1}^{\infty}$ where $d = 2$ for representation **R1** and d is the least common denominator of the $\tilde{b}_j$, $j \in \mathbf{s}$, for representation **R2**. Then, the family of approximating pairs $(\mathbf{CP}_N, \theta_N)$, $N \in \mathbf{N}$, constitute consistent approximations for the pair $(\mathbf{CP}, \theta)$. □

6. NUMERICAL RESULTS

We will now illustrate the usefulness of our analysis with an example. First, recall that by (5.5a), for $\eta = (\xi, u) \in H_N$ and $\nu \in \mathbf{q}$, $\nabla_u f_N^\nu(\eta) = V_{\mathbf{A},N}^{-1}(\bar{\gamma}_u^\nu(\bar{\eta})\mathbf{M}_N^{-1})$, where $\bar{\gamma}_u^\nu(\bar{\eta})$, defined in (5.5c), is the gradient of $\bar{f}_N^\nu(\cdot)$ with respect to the standard inner product on $\mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$. The gradient of $\bar{f}_N^\nu(\cdot)$ with respect to the inner product on $\bar{L}_N$ is given by $\nabla_u \bar{f}_N^\nu(\bar{\eta}) \triangleq V_{\mathbf{A},N}(\nabla_u f_N^\nu(\eta)) = \bar{\gamma}_u^\nu(\bar{\eta})\mathbf{M}_N^{-1}$, and satisfies

$$\langle \nabla_u f_N^\nu(\eta), \delta u \rangle_2 = \langle \nabla_u \bar{f}_N^\nu(\bar{\eta}), \delta\bar{u} \rangle_{\bar{L}_N} = \langle \bar{\gamma}_u^\nu(\bar{\eta}), \delta\bar{u} \rangle_{l_2} . \tag{6.1}$$

where $\bar{\eta} = W_{\mathbf{A},N}(\eta)$ and $\delta\bar{u} = V_{\mathbf{A},N}(\delta u)$. It is well-known that a change in the inner product on a space is equivalent to a transformation of coordinates. Since existing optimization software uses the standard inner-product, it is convenient to define a transformation of coordinates that results in a gradient which satisfies (6.1) for the standard l_2 inner product on $\mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$.

To accomplish this, let $\tilde{L}_N = \mathbb{R}^N \times \mathbb{R}^r \times \mathbb{R}^m$ endowed with the standard (Euclidean) inner product, and let the transformation $Q : \bar{L}_N \to \tilde{L}_N$, be defined by

$$\tilde{u} = Q(\bar{u}) = \bar{u}\,\mathbf{M}_N^{1/2} , \tag{6.2a}$$

where $\mathbf{M}_N$ is defined in (5.4). For each $\nu \in \mathbf{q}$, let $\tilde{f}_N^\nu : (\mathbb{R}_N \times \tilde{L}_N) \to \mathbb{R}$ be defined by

$$\tilde{f}_N^\nu((\xi, \tilde{u})) \triangleq \bar{f}_N^\nu((\xi, Q^{-1}\tilde{u})) = \bar{f}_N^\nu((\xi, \bar{u})) . \tag{6.2b}$$

Let $\bar{\eta} = W_{\mathbf{A},N}(\eta) = (\xi, \bar{u})$ and $\tilde{\eta} = (\xi, Q(\bar{u}))$. Then, using the chain rule,

$$\nabla_v \tilde{f}_N^\nu(\tilde{\eta}) = Q^{-1}\left(\frac{d}{d\bar{u}}\bar{f}_N^\nu(\bar{\eta})\right) = \bar{\gamma}_u^\nu(\bar{\eta})\,\mathbf{M}_N^{-1/2} . \tag{6.2c}$$

Thus, with $\delta\bar{u} = W_{\mathbf{A},N}(\delta u)$, $\langle \nabla_v \tilde{f}_N^\nu(\tilde{\eta}), Q(\delta\bar{u}) \rangle_{l_2} = \langle \nabla_u \bar{f}_N^\nu(\bar{\eta}), \delta\bar{u} \rangle_{\bar{L}_N} = \langle \nabla_u f_N^\nu(\eta), \delta u \rangle_2$. Implicitly, the transformation, Q, creates an orthonormal basis for L_N. With this transformation, the approximating problems can be solved using the standard norm and inner products on Euclidean space for which any of the standard nonlinear programming methods apply directly. It is important to note, however, that control constraints are also transformed. Thus, the constraint $\bar{u} \in \bar{\mathbf{U}}$ becomes $\tilde{u}\,\mathbf{M}_N^{-1/2} \in \bar{\mathbf{U}}$. For representation **R1**, $\mathbf{M}_N^{-1/2}$ is not diagonal (except if $r = 1$). This means that the transformed control constraints will, for each k, involve linear combinations of the control samples $\tilde{u}_k^j$, $j \in \mathbf{r}$.

We will now present a numerical example which shows, in particular, that without using the above transformation, the approximating problems can be ill-conditioned.

Example. Consider the following linear-quadratic problem taken from [16]:

$$\min_{u \in \mathbf{U}} f(u)\,, \quad f(u) \triangleq x_2^u(1)\,, \tag{6.4a}$$

where $x(t) = (x_1(t), x_2(t))^T$ and

$$\dot{x} = \begin{bmatrix} 0.5x_1 + u \\ 0.625x_1^2 + 0.5x_1u + 0.5u^2 \end{bmatrix}\,, \quad x(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}\,, \quad t \in [0,1]\,. \tag{6.4b}$$

The solution to this problem is given by

$$u^*(t) = -(\tanh(1-t) + 0.5)\cosh(1-t)/\cosh(1)\,, \quad t \in [0,1]\,, \tag{6.5}$$

with optimal cost $x_2^*(1) = e^2 \sinh(2)/(1+e^2)^2 \approx 0.380797$.

The approximating cost function is $f_N(u) = (\,0\ 1\,)\,\bar{x}_N^u$ where $\{\bar{x}_k^u\}_{k=0}^N$ is the solution of the approximating problem for a given control $u \in L_N$. We discretized this problem using two common RK methods whose Butcher arrays are:

$$\mathbf{A}_1 \;=\; \begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 1 & -1 & 2 & \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \qquad\qquad \mathbf{A}_2 \;=\; \begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

The scaling matrices $\mathbf{M}_N$ used to define the transformation Q in (6.2a) are given by (5.4) with

$$M = M_1 = \frac{1}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix}\,, \qquad M = M_2 = \frac{1}{6}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}\,, \tag{6.6}$$

which are the same for both RK methods since in $\mathbf{A}_2$, $c_2 = c_3 = 1/2$ implies $r = 3$ and $\tilde{b}_2 = 2/3$.

We solved the approximating problems with the initial guess $u(t) = 0$, $t \in [0,1]$, stopping when the stepsize was below machine precision (2.22e-16) or the norm of the gradient was smaller than the square root of machine precision. Table 1 shows the number of iterations required to solve the approximating problems for different discretization levels N with and without the transformation (6.2b). We see that solving the discretized problems without the transformation required about 200% more iterations than with the transformation. The situation can be much worse for other RK methods. The choice of representation **R1** versus representation **R2** had no effect on the number of iterations required.

	Number of Iterations	
N	$M = M_i\ ,\ i = 1,2$	$M = \frac{1}{N} I$
10	17	52
20	18	52
40	17	52
80	15	50

Table 1: Conditioning Effect of the Transformation Q on Approximating Problems.

We use the second RK method primarily to demonstrate the advantage of treating the samples arising from repeated c_i values in the Butcher array as the same sample (see Remark 4.7). Let $\{ u_N^* \}_{N \in \mathbf{N}}$, where $\mathbf{N} \subset \mathbb{N}$, be solutions of $\mathbf{CP}_N$ and suppose $u_N^* \to u^*$ where u^* is a solution of **CP**. In [16, Thm. 3.1], Hager establishes, for symmetric RK methods [1,28], a tight upper bound on the error $E_N^u \triangleq \| V_{\mathbf{A},N}(u^*) - V_{\mathbf{A},N}(\bar{u}_N^*) \|_\infty$, of second order in $\Delta = 1/N$ for smooth, unconstrained problems. Note that $V_{\mathbf{A},N}(u^*)_k^j = u^*(\tau_{k,j})$, $k \in \mathcal{N}$ and $j \in \mathbf{r}$ because $u^*(\cdot)$ is smooth for smooth problems [26]. Hager used the problem given in (6.4a) to demonstrate the tightness of this bound. For the particular RK method described with the Butcher array $\mathbf{A}_2$, we can state the following improved result:

Proposition 6.1. Let $\mathbf{CP} \triangleq \min_{u \in \mathbf{U}} f(x^u(1))$, u unconstrained, and suppose that $f(\cdot)$ and $h(\cdot,\cdot)$ in (3.1) are four times continuously differentiable. Suppose the approximating problems $\mathbf{CP}_N$ are produced by discretizing **CP** with the fourth order RK method with Buthcer array $\mathbf{A}_2$: $c = \{ 0, 1/2, 1/2, 1 \}$, $b = \{ 1/6, 1/3, 1/3, 1/6 \}$ and the non-zero entries of A are $a_{2,1} = a_{3,2} = 1/2$ and $a_{4,3} = 1$. Let $\{ u_N^* \}_{N \in \mathbf{N}}$, where $\mathbf{N} \subset \mathbb{N}$, be solutions of $\mathbf{CP}_N$ and suppose $u_N^* \to u^*$ where u^* is a solution of **CP**. Then $E_N^u \triangleq \| V_{\mathbf{A},N}(u^*) - V_{\mathbf{A},N}(\bar{u}_N^*) \|_\infty = O(\Delta^3)$.

Sketch of Proof. In [16], it is shown, using a reasonable non-singularity assumption on the Hessians of $f_N(\cdot)$, that the accuracy of the solutions of the approximating problems is determined by the size of the discrete-time gradient of the approximating problem at $\bar{u}^* \triangleq V_{\mathbf{A},N}(u^*)$, that is, $\| \bar{\gamma}_u(\bar{u}^*) \|$. This, in turn, is a function of the accuracy of the state and adjoint approximations. For the RK method under consideration, Hager shows that the variables $\bar{u}^{*j}_k$, $k \in \mathcal{N}$ and $j = 1,3$ are third order approximations to $u^*(t_k)$ and $u^*(t_k + \Delta)$, respectively. Thus, we need only show that u_k^2 is a third order approximation to $u^*(t_k + \Delta/2)$.

Let $Y_{k,1} = \bar{x}_k + \Delta/2 h(\bar{x}_k, \bar{u}_k^1)$ and $Y_{k,2} = \bar{x}_k + \Delta/2 h(\bar{x}_k + \Delta/2 h(\bar{x}_k, \bar{u}_k^1), \bar{u}_k^2)$ represent the second and intermediate values used by the RK method at the k-th time-step with $\bar{u} = \bar{u}^*$. Hager introduces a clever transformation, specific to symmetric RK methods, for the adjoint variables so that they can be viewed as being calculated with the same RK method used to compute the state variables, but run backwards in time. The intermediate adjoint variables of interest here are denoted by $q(2,k)$ and $q(1,k)$. With this transformation, the discrete-time gradients for the approximating problems has the same form as the continuous-time gradient for the original problem. Since $c_2 = c_3 = 1/2$, $\bar{\gamma}_u(\bar{u}^*)_k^2 = 2\Delta/3[1/2h_u(Y_{k,1}, \vec{u}^{*2}_k)^T q(1,k) + 1/2h_u(Y_{k,2}, \vec{u}^{*2}_k)^T q(2,k)]$. Since $2\Delta/3h_u(x^{u^*}(t_k + \Delta/2), u^*(t_k + \Delta/2))^T p^{u^*}(t_k + \Delta/2) = 0$, the size of $\bar{\gamma}_u(\bar{u}^*)_k^2$ is the maximum of $\|(Y_{k,1} + Y_{k,2})/2 - x^{u^*}(t_k + \Delta/2)\|$ and $\|(q(2,k) + q(1,k))/2 - p^{u^*}(t_k + \Delta/2)\|$. First,

$$w(k) \triangleq \frac{Y_{k,1} + Y_{k,2}}{2} = x_k + \frac{\Delta}{4}[h(x_k, u_k^1) + h(x_k + \frac{\Delta}{2}h(x_k, u_k^1), u_k^2)]$$

$$= x_k + \frac{\Delta'}{2}[h(x_k, u_k^1) + h(x_k + \Delta' h(x_k, u_k^1), u_k^2)] , \tag{6.7}$$

where $\Delta' = \Delta/2$. Thus, w(k) is produced by the modified Euler rule applied to x_k. Since the local truncation error for the modified Euler rule is order $O(\Delta^3)$ and x_k is order $O(\Delta^4)$, $\|w(k) - x^{u^*}(t_k + \Delta/2)\|$ is order $O(\Delta^3)$. In the same way, it can be shown that $\|q(2,k) + q(1,k))/2 - p^{u^*}(t_k + \Delta/2)\|$ is $O(\Delta^3)$. Thus, we can conclude that $\|\bar{\gamma}_u(\vec{u}^*)_k^2\| = O(\Delta^3)$ for all $k \in \mathcal{N}$. This implies that the solutions of the approximating problems satisfy $\|\vec{u}^{*j}_{N,k} - u^*(\tau_{k,j})\| = O(\Delta^3)$ for all $k \in \mathcal{N}$ and $j \in \mathbf{r}$. □

	Accuracy of Solutions				Number of Iterations	
N	E_N^u	E_N^u / E_{2N}^u	E_N^f	E_N^f / E_{2N}^f	$M = M_i\ ,\ i = 1,2$	$M = \frac{1}{N} I$
10	1.48e-4		2.86e-7		16	20
20	1.87e-5	7.94	1.76e-8	16.22	15	20
40	2.34e-6	7.99	1.09e-9	16.13	15	23
80	3.07e-7	7.62	6.80e-11	16.07	15	27

Table 2: Rate of Convergence; Conditioning Effect of the Transformation Q.

Table 2 summarizes our numerical results using the RK method with Butcher array $\mathbf{A}_2$. The first column gives the discretization level. Columns 2 and 3 show that doubling the discretization results in an eight-fold reduction in the control error. Thus, as predicted by Proposition 6.1, E_N^u is $O(\Delta^3)$. The next two columns, agreeing with Hager's observations that the optimal trajectories of the approximating problem converge to those of the original problem with the same order as the order of the symmetric RK method, show that $E_N^f \triangleq |f(u^*) - f_N(\bar{u}_N^*)|$, is order $O(\Delta^4)$. Finally, we include in the last two columns the number of iterations required to solve the approximate problem with and without the transformation Q. The effect of scaling is less spectacular than in the previous method, but still significant. The untransformed problem takes 25% to 80% more iterations to solve than the transformed problem.

The last table shows the accuracy of the approximating gradients for the particular control $u(t) = -1 + 2t$. The first column shows the discretization level N. The second and third columns confirm that the gradients for the approximating problems converge to the gradients of the original problem. Note that, based on the proof of Theorem 5.6, it is enough to show that the gradients converge at the points τ_{k,i_j}, $k \in \mathcal{N}$, $j \in \mathbf{r}$, and $i_j \in I$. The fourth column of Table 1 shows that the gradients that would result if one treated the discrete-time optimal control problem directly do not converge.

	$M = M_1$	$M = M_2$	$M = \frac{1}{N} I$
N	$\lVert V_{\mathrm{A},N}(\nabla f(u)) - \bar{\gamma}_u \mathbf{M}_N^{-1} \rVert_\infty$	$\lVert V_{\mathrm{A},N}(\nabla f(u)) - \bar{\gamma}_u \mathbf{M}_N^{-1} \rVert_\infty$	$\lVert V_{\mathrm{A},N}(\nabla f(u)) - N \bar{\gamma}_u \rVert_\infty$
10	1.67e-3	6.46e-4	1.48
20	3.77e-4	8.31e-5	1.48
40	9.94e-5	1.05e-5	1.48
80	2.55e-5	1.33e-6	1.48

Table 3: Convergence of Gradients.

7. CONCLUSION

We have shown that a large class of Runge-Kutta integration methods can be used to construct consistent approximations to continuous time optimal control problems. The construction of consistent approximations is not unique: it is determined by the selection of families of finite dimensional subspaces of the control space. When the elements of these subspaces are discontinuous functions, appropriate extensions of Runge-Kutta methods must be used. However, in this case, not all Runge-Kutta methods can be used because some Runge-Kutta methods do not result in consistent approximations. This was observed both numerically and by failure to prove consistency of approximation with these methods. We have considered two selections of control subspaces in this paper, one defined by piecewise polynomial functions and one by piecewise constant functions. Splines can also be used and are treated in Appendix A2. Each selection has some advantages and some disadvantages. A final selection has to be made on the basis of secondary considerations, such as the importance of approximate solutions satisfying the original control constraints, the form that the control constraints take in the discrete-time optimal control problems, or the accuracy with which the differential equation is integrated.

As in our case, the basis functions that are used implicitly to define the finite dimensional control subspaces may turn out to be non-orthonormal. In this case care must be taken to introduce a non-

Euclidean inner product and corresponding norm in solving the resulting approximating discrete time optimal control problems. Neglecting to do so amounts to a change of coordinates that can lead to serious ill-conditioning. This ill-conditioning is demonstrated in Section 6.

Finally, the use of the framework of consistent approximations opens up the possibility of developing optimal discretization strategies, such as those considered for semi-infinite programming in [17]. Such a strategy provides rules for selecting the number of approximating problems to be used as well as the discretization level, the order of the RK method, and the number of iterations of a particular optimization algorithm to be applied for each such approximating problem, so as to minimize the computing time needed to reach a specified degree of accuracy in solving an optimal control problem. We hope to develop such results in the near future.

APPENDIX A1

In this Appendix we will collect a few results used in the analysis of Sections 4 and 5. We will continue to use the notation of Section 4: $\Delta = 1/N$, $t_k = k\Delta$, and $\tau_{k,i} = t_k + c_i\Delta$.

Lemma A1.1. For representation **R1**, suppose that Assumptions 3.1*(a)* and 4.1 hold. For representation **R2**, suppose that Assumptions 3.1*(a)*, 4.1, and 4.5 hold. For any bounded subset $S \subset \mathbf{B}$, there exists a $\kappa < \infty$ such that for any $\eta = (\xi, u) \in S \cap H_N$, $\|\delta_k\| \leq \kappa\Delta^2$ for all $k \in \mathcal{N}$, where

$$\delta_k \triangleq x^{\eta}(t_k) - x^{\eta}(t_{k+1}) + \Delta\sum_{i=1}^{s} b_i h(x^{\eta}(t_k), u[\tau_{k,i}]), \quad k \in \mathcal{N}, \tag{A1.1}$$

$x^{\eta}(\cdot)$ is the solution of the differential equation (3.1) and $u[\tau_{k,i}]$ is defined by (4.2).

Proof: Let $\tilde{b}_j$ and d_j be as defined in (4.10) and, for $j \in \mathbf{r}$, let $i_j \in I$ where I is given by (4.4a). Then, writing $x(\cdot) = x^{\eta}(\cdot)$, since the solution of (3.1) satisfies $x(t_{k+1}) = x(t_k) + \int_{t_k}^{t_{k+1}} h(x(t), u(t))\, dt$, we see that

$$\delta_k = \Delta\sum_{i=1}^{s} b_i h(x(t_k), u[\tau_{k,i}]) - \int_{t_k}^{t_{k+1}} h(x(t), u(t))\, dt$$

$$= \sum_{j=1}^{r}\int_{t_k+d_{j-1}}^{t_k+d_j} h(x(t_k), u[\tau_{k,i_j}])\, dt - \sum_{j=1}^{r}\int_{t_k+d_{j-1}}^{t_k+d_j} h(x(t), u(t))\, dt\ , \tag{A1.2a}$$

because $d_j - d_{j-1} = \Delta\tilde{b}_j$, $u[\tau_{k,i_j}] = u[\tau_{k,i}]$ for all $i \in I_j$, $d_0 = 0$ and, by Assumption 4.1*(b')*, $d_r = \Delta\sum_{j=1}^{r}\tilde{b}_j = \Delta\sum_{j=1}^{s} b_j = \Delta$. Thus,

$$\|\delta_k\| \leq \sum_{j=1}^{r}\int_{t_k+d_{j-1}}^{t_k+d_j} \|h(x(t_k), u[\tau_{k,i_j}]) - h(x(t), u(t))\|\, dt$$

$$\le \sum_{j=1}^{r} \int_{t_k+d_{j-1}}^{t_k+d_j} \kappa_1[\|x(t_k)-x(t)\| + \|u[\tau_{k,i_j}] - u(t)\|]\,dt \ , \tag{A1.2b}$$

where $\kappa_1 < \infty$ is as in Assumption 3.1*(a)* and $d_j - d_{j-1} > 0$ by Assumption 4.1*(b')*. Now, for $t \in [t_k, t_{k+1}]$, there exists $\kappa_2 < \infty$ such that

$$\|x(t_k) - x(t)\| \le \int_{t_k}^{t} \|h(x(t), u(t))\|\,dt \le \int_{t_k}^{t_{k+1}} \kappa_2[\|x(t)\| + 1]\,dt \tag{A1.3}$$

by Assumption 3.1*(a)* and the fact that S is bounded. Also because S is bounded, if follows from Theorem 3.2*(ii)* that there exists $L < \infty$ such that $\|x(t)\| \le \kappa_3[\|\xi\| + 1] \le L$. Thus, for $t \in [t_k, t_{k+1}]$, $\|x(t_k) - x(t)\| \le \int_{t_k}^{t_{k+1}} \kappa_2[L+1]\,dt = \Delta\kappa_2(L+1)$. There also exists $\kappa_4 < \infty$ such that, for any $k \in \mathcal{N}$ and $j \in \mathbf{r}$, $\|u[\tau_{k,i_j}] - u(t)\| \le \kappa_4\Delta$ for $t \in [t_k + d_{j-1}, t_k + d_j)$ since *(i)* for representation **R1** $\tau_{k,i_j} \in [t_k + d_{j-1}, t_k + d_j)$ by Assumption 4.1*(a)*, $0 \le d_j \le \Delta$ for $j = 0, \ldots, r$ by Assumption 4.1*(b')* and $\bar{u} \in L_N^1$ is a polynomial on $[t_k, t_{k+1})$ and hence Lipschitz continuous on each sub-interval $[t_k + d_{j-1}, t_k + d_j)$; and *(ii)* for representation **R2**, $\bar{u} \in L_N^2$ is constant on $t \in [t_k + d_{j-1}, t_k + d_j)$ and $\tau_{k,i_j} \in [t_k + d_{j-1}, t_k + d_j]$ by Assumption 4.5. Therefore,

$$\|\delta_k\| \le \sum_{j=1}^{r} \int_{t_k+d_{j-1}}^{t_k+d_j} \kappa_1(\kappa_2(L+1) + \kappa_4)\Delta\,dt = \kappa\Delta \sum_{j=1}^{r} \int_{t_k+d_{j-1}}^{t_k+d_j} dt = \kappa\Delta^2 , \tag{A1.4}$$

where $\kappa = \kappa_1(\kappa_2(L+1) + \kappa_4)$. This completes our proof. □

The next lemma concerns the functions $K_{k,i} = K_i(\bar{x}_k, \omega_k)$ of the RK method defined by (4.3a,b). The proof of this result is easily obtained from the proof for Lemma 222A in [7, p. 131].

Lemma A1.2 Suppose Assumptions 3.1*(a)* holds. Let $S \subset \mathbf{B}$ be bounded. Then there exists $L < \infty$ and $N^* < \infty$ such that for all $N \ge N^*$, $\eta \in S \cap H_N$, $k \in \mathcal{N}$ and $i \in \mathbf{s}$,

$$\|K_{k,i} - h(\bar{x}_k, u[\tau_{k,i}])\| \le L\,\Delta\,. \tag{A1.5}$$

□

Next, we present a proof of Lemma 4.8.

Proof of Lemma 4.8.

(i) Convergence Let $\eta = (\xi, u) \in S \cap H_N$ and, for $k \in \mathcal{N}$ let $e_k \triangleq \bar{x}_k^{\eta} - x^{\eta}(t_k)$. Then, $\|e_0\| = 0 \le \kappa\Delta$ and by adding and subtracting terms,

$$e_{k+1} = \bar{x}_k^{\eta} + \Delta\sum_{i=1}^{s} b_i K_{k,i} - x^{\eta}(t_{k+1})$$

$$= e_k + \left[x^{\eta}(t_k) - x^{\eta}(t_{k+1}) + \Delta\sum_{i=1}^{s} b_i h(x^{\eta}(t_k), u[\tau_{k,i}]) \right] + \Delta\sum_{i=1}^{s} b_i \left[K_{k,i} - h(x^{\eta}(t_k), u[\tau_{k,i}]) \right] \ . \tag{A1.6}$$

The norm of the second term in this expression is bounded by $\kappa_1\Delta^2$ by Lemma A1.1 where $\kappa_1 < \infty$. Using

Lemma A1.2, Assumption 3.1*(a)*, and the fact that $|b_i| \leq 1$ by Assumption 4.1*(b)*, we conclude for the third term that, there exists $\kappa_2 < \infty$ such that

$$\Delta \Big\| \sum_{i=1}^{s} b_i \Big[K_{k,i} - h(x^{\eta}(t_k), u[\tau_{k,i}]) \Big] \Big\|$$

$$\leq \Delta \sum_{i=1}^{s} \| K_{k,i} - h(\bar{x}_k^{\eta}, u[\tau_{k,i}]) \| + \Delta \sum_{i=1}^{s} \| h(\bar{x}_k^{\eta}, u[\tau_{k,i}]) - h(x^{\eta}(t_k), u[\tau_{k,i}]) \|$$

$$\leq \Delta^2 Ls + \Delta \kappa_2 s \| e_k \| \,. \tag{A1.7}$$

Thus, for all $k \in \mathcal{N}$

$$\| e_{k+1} \| \leq (1 + \kappa_2 \Delta s) \| e_k \| + \kappa_3 \Delta^2 \,. \tag{A1.8}$$

where $\kappa_3 = \kappa_1 + Ls$. Solving (A1.8), we see that for all $k \in \mathcal{N}$ $\| e_k \| \leq (1 + \kappa_2 \Delta s)^N \| e_0 \| + \kappa_3' \Delta \leq \kappa \Delta$. This proves (4.18a).

We prove (4.18b) in two steps. First suppose that $H_N = H_N^1 = \mathbb{R}^{n_x} \times L_N^1$ and let $\eta_1 \in S \cap H_N^1$ be given. The expansion based on higher-order derivatives (see [7]) needed to prove (4.18b) requires smoothness of $h(x,u)$ between time steps. By the way we have defined control samples $u[\tau_{k,i}]$, the samples of $u_1 \in L_N^1$ used by the RK method correspond to polynomials between time steps, implying that $h(x(t), u_1(t))$ is smooth between time-steps. Using the same type of reasoning in the proof of Lemma A1.1, we conclude that there exists $\kappa < \infty$, independent of η, such that (4.18b) holds for representation **R1**. Next, to prove (4.18b) for representation **R2**, let $H_N = H_N^2 = \mathbb{R}^{n_x} \times L_N^2$. Let $\eta_2 = (\xi, u_2) \in S \cap H_N^2$ be given and let $\eta_1 = (\xi, u_1) = (W_{A,N}^1)^{-1}(W_{A,N}^2(\eta_2)) \in H_N^1$ so that $V_{A,N}^1(u_1) = V_{A,N}^2(u_2)$. Then for any $t \in [0,1]$,

$$\| x^{\eta_1}(t) - x^{\eta_2}(t) \| = \Big\| \int_0^t h(x^{\eta_1}(s), u_1(s)) - h(x^{\eta_2}(s), u_2(s)) ds \Big\|$$

$$\leq \Big\| \int_0^t \tilde{h}(x^{\eta_1}(s)) - \tilde{h}(x^{\eta_2}(s)) + B(u_1(s) - u_2(s)) ds \Big\|$$

$$\leq \kappa_1 \int_0^t \| x^{\eta_1}(s) - x^{\eta_2}(s) \| ds + \Big\| \int_0^t B(u_1(s) - u_2(s)) ds \Big\| \,, \tag{A1.9a}$$

by Assumption 3.1*(a)*. Using the Bellman-Gronwall lemma, we conclude that for any $t \in [0,1]$,

$$\| x^{\eta_1}(t) - x^{\eta_2}(t) \| \leq \kappa_1 e^{\kappa_1} \| B \| \Big\| \int_0^t (u_1(s) - u_2(s)) ds \Big\| \,. \tag{A1.9b}$$

Now, let $\dot{z}^1(t) \triangleq u_1(s)$, $t \in [0,1]$, $z^1(0) = \xi$ and $\dot{z}^2(t) \triangleq u_2(s)$, $t \in [0,1]$, $z^2(0) = \xi$. Let $\bar{z}_k^1$ and $\bar{z}_k^2$, $k \in \mathcal{N}$ be the computed solution of $z^1(t)$ and $z^2(t)$, respectively, using the RK method under consideration. We note that $\bar{z}_k^1 = \bar{z}_k^2$ for all $k \in \mathcal{N}$ since $V_{A,N}^1(u_1) = V_{A,N}^2(u_2)$. Then, since (4.18b) holds for representation **R1**, $\bar{z}_k^1 = z^1(t_k) + \delta_k$ where $\| \delta_k \| \leq \kappa_2 / N^p$, $\kappa_2 < \infty$, for all $k \in \mathcal{N}$. Also, from (4.3a,b),

$$\bar{z}_{k+1}^2 = \bar{z}_k^2 + \sum_{i=1}^{s} b_i u_2[\tau_{k,i}] = \bar{z}_k^2 + \sum_{j=1}^{r} \int_{t_k+d_{j-1}}^{t_k+d_j} u_2(s)ds = z^2(t_{k+1}) , \tag{A1.9c}$$

since $\tau_{k,j} \in [t_k+d_{j-1}, t_k+d_j)$ (by Assumption 4.5) with $u_2(\cdot)$ constant on these intervals, and $d_r = \Delta$ by Assumption 4.1*(b')*. Since $\bar{z}_k^1 = \bar{z}_k^2$, we must have

$$z^1(t_k) - z^2(t_k) = \bar{z}_k^1 + \delta_k - \bar{z}_k^2 = \delta_k , \quad \forall k \in \mathcal{N} . \tag{A1.9d}$$

Hence, we conclude that

$$\| \int_0^{t_k} (u_1(s) - u_2(s))ds \| = \| z^1(t_k) - z^2(t_k) \| = \| \delta_k \| \le \kappa_2 / N^{\rho} . \tag{A1.9e}$$

Therefore,

$$\| x^{\eta_2}(t_k) - \bar{x}_k^{\eta_2} \| \le \| x^{\eta_2}(t_k) - x^{\eta_1}(t_k) \| + \| x^{\eta_1}(t_k) - \bar{x}_k^{\eta_1} \| + \| \bar{x}_k^{\eta_1} - \bar{x}_k^{\eta_2} \| \le \kappa' / N^{\rho} , \quad \forall k \in \mathcal{N} , \tag{A1.9f}$$

where we have used (A.19b) and (A.19e), the fact that $\| x^{\eta_1}(t_k) - \bar{x}_k^{\eta_1} \| \le \kappa_2 / N^{\rho}$ since (4.18b) holds for $\eta_1 \in S \cap H_N^1$ by the first part of this discussion and the fact that $\bar{x}_k^{\eta_1} = \bar{x}_k^{\eta_2}$ since $u_1[\tau_{k,i}] = u_2[\tau_{k,i}]$. Thus (4.18b) holds for representation **R2** under the stated conditions.

(ii) Lipschitz Continuity First observe that the term $\sum_{i=1}^{s} b_i K_{k,i}$ in equation (4.1a) is Lipschitz continuous on bounded sets with respect to $\bar{x}$ and $\bar{u}$, with constant κ_Δ, since it is a finite composition of Lipschitz continuous functions. The constant κ_Δ decreases monotonically to the Lipschitz constant κ of $h(\cdot,\cdot)$ as $\Delta \to 0$. Thus, there is a single Lipschitz constant κ_1 for $\sum_{i=1}^{s} b_i K_{k,i}$ which is good for any Δ.

Let $\bar{\eta} = (\xi, \bar{u}) = W_{\mathbf{A},N}(\eta)$ and $\bar{\eta}' = (\xi, \bar{u}') = W_{\mathbf{A},N}(\eta')$. Define $\delta_k \triangleq \bar{x}_k^{\eta} - \bar{x}_k^{\eta'}$ and $\delta\bar{u}_k = \bar{u}_k - \bar{u}_k'$. Then, $\delta_0 = \xi - \xi'$, and

$$\delta_{k+1} = \delta_k + \Delta \sum_{i=1}^{s} b_i [K_{k,i} - K_{k,i}'] , \quad k \in \mathcal{N} . \tag{A1.10}$$

Taking norms and using the assumption that $| b_i | \le 1$, we obtain

$$\begin{aligned} \| \delta_{k+1} \| &\le \| \delta_k \| + \Delta \sum_{i=1}^{s} | b_i | \| K_{k,i} - K_{k,i}' \| \\ &\le \| \delta_k \| + \Delta \kappa_1 \sum_{i=1}^{s} [\| \delta_k \| + \| \delta \bar{u}_k \|] \\ &\le (1 + \Delta \kappa_1 s) \| \delta_k \| + \Delta \kappa_1 s \left[\sum_{j=1}^{r} \| \delta \bar{u}_k^j \|^2 \right]^{1/2} \\ &\le (1 + \Delta \kappa_1 s) \| \delta_k \| + \Delta \kappa_2 s \,\text{trace}\big((\delta \bar{u}_k^1 \cdots \delta \bar{u}_k^r) M \, (\delta \bar{u}_k^1 \cdots \delta \bar{u}_k^r)^T \big)^{1/2} , \end{aligned} \tag{A1.11}$$

where $M = M_1$ for representation **R1** and $M = M_2$ for representation **R2** and $\kappa_2 = \kappa_1 / \lambda_{\min}(M) < \infty$. Now we define $\varepsilon_0 \triangleq \|\delta_0\|$ and, for $k \in \mathcal{N}$,

$$\varepsilon_{k+1} \triangleq (1 + \Delta\kappa_1 s)\varepsilon_k + \Delta\kappa_2 s \,\text{trace}\big((\delta\bar{u}_k^1 \cdots \delta\bar{u}_k^r) M\, (\delta\bar{u}_k^1 \cdots \delta\bar{u}_k^r)^T\big)^{1/2}, \tag{A1.12a}$$

so that $\|\delta_k\| \leq \varepsilon_k$. Solving (A1.12a), we obtain that

$$\varepsilon_k = (1 + \Delta\kappa_1 s)^k \varepsilon_0 + \Delta\kappa_2 s \sum_{j=0}^{k-1} (1 + \Delta\kappa_1 s)^j \,\text{trace}\big((\delta\bar{u}_j^1 \cdots \delta\bar{u}_j^r) M\, (\delta\bar{u}_j^1 \cdots \delta\bar{u}_j^r)^T\big)^{1/2}, \tag{A1.12b}$$

Therefore, assuming without loss of generality that $\kappa_2 \geq 1$, we have, for all $k \in \mathcal{N}$,

$$\|\delta_k\| \leq (1 + \Delta\kappa_1 s)^N \kappa_2 s \left[\|\delta_0\| + \Delta \sum_{j=1}^{N-1} \text{trace}\big((\delta\bar{u}_j^1 \cdots \delta\bar{u}_j^r) M\, (\delta\bar{u}_j^1 \cdots \delta\bar{u}_j^r)^T\big)^{1/2} \right]$$

$$\leq L \left[\|\xi - \xi'\|^2 + \Big[\Delta \sum_{j=0}^{N-1} \text{trace}\big((\delta\bar{u}_j^1 \cdots \delta\bar{u}_j^r) M\, (\delta\bar{u}_j^1 \cdots \delta\bar{u}_j^r)^T\big)\Big]^2 \right]^{1/2}$$

$$= L \left[\|\xi - \xi'\|^2 + \|\bar{u} - \bar{u}'\|^2 \right]^{1/2} = L \|\bar{\eta} - \bar{\eta}'\|_{\bar{H}_N}, \tag{A1.12c}$$

where $L = (N + 1)^{1/2}(1 + \Delta\kappa_1 s)^N \kappa_2 s$ and we have made use of the fact that if $a_j \geq 0$ for $j \in \mathbf{q}$ then $\sum_{j=1}^{q} a_j \leq q^{1/2}(\sum_{j=1}^{q} a_j^2)^{1/2}$. □

Lemma A1.3. Suppose that Assumptions 3.1, 4.1 hold for representation **R1** and that Assumptions 3.1, 4.1, and 4.5 hold for representation **R2**. For any $S \subset \mathbf{B}$ bounded, there exists $\kappa < \infty$ and $N^* < \infty$ such that for any $\eta \in S \cap H_N$ and $N \geq N^*$,

$$\|\bar{p}_k - p(t_k)\| \leq \frac{\kappa}{N}, \quad k \in \{0, \ldots, N\}, \tag{A.13}$$

where $p(\cdot)$ is the solution to the adjoint differential equation (3.6c) and $\{\bar{p}_k\}_{k=0}^{N}$ is the solution to the corresponding adjoint difference equation (5.5d).

Proof. Proceeding as in the proof of Lemma 4.8*(i)*, if we define $e_{k+1} \triangleq \bar{p}_{k+1} - p(t_{k+1})$ we can show that

$$\|e_k\| \leq L_1 \|e_{k+1}\| + L_2 \Delta^2, \quad k \in \mathcal{N}, \tag{A1.14}$$

where $L_1, L_2 < \infty$, using *(i)* the fact that

$$\bar{p}_k = F_x(\bar{x}_k, \bar{u}_k)^T \bar{p}_{k+1} = \bar{p}_{k+1} + \Delta \sum_{i=1}^{s} b_i h_x(\bar{x}_k, u[\tau_{k,i}])^T \bar{p}_{k+1} + O(\Delta^2), \tag{A1.15}$$

(ii) Lemma A1.1 with $h(x(t_k), u[\tau_{k,i}])$ replaced by $-h_x(x(t_k), u[\tau_{k,i}])^T p(t_{k+1})$ and *(iii)* the result of Lemma 4.8*(i)* that $\|x(t_k) - \bar{x}_k\| \leq \kappa\Delta$ for all $k \in \mathcal{N}$. Now, by Assumption 3.1*(b)* and Lemma 4.8*(i)*, there exists $\kappa_1 < \infty$ such that

$$\|e_N\| = \|\bar{p}_N - p(1)\| \leq \|\zeta_x(\xi, \bar{x}_N)^T - \zeta_x(\xi, x(1))^T\| \leq \kappa_1 \|\bar{x}_N - x(1)\| \leq \kappa_2 \Delta \,, \tag{A1.16}$$

where $\kappa_2 = \kappa\kappa_1$. Thus, solving (A1.14) we conclude that for all $k \in \mathcal{N}$,

$$\|e_k\| \leq (L_1)^N \|e_N\| + L_2' \Delta \,,$$

which, with (A1.16), proves (A1.13). □

APPENDIX A2

In this appendix, we use splines as the finite dimensional control elements in the construction of approximating problems, for optimal control problems with enpoint inequality constraints and box-type control constraints. We show that the resulting approximating problems, along with their optimality functions, are consistent approximations to the original problem with its optimality function. In the process, we will develop some results for splines that are interesting for their own sake.

We will construct our finite dimensional control spaces using spline basis functions, *i.e.* B-splines [4]. Thus, for $r \in \mathbb{N}$, $r \geq 1$, let

$$L_N^{(r)} \triangleq \{ u \in L_2^m[0,1] \mid u(t) = \sum_{k=1}^{N+r-1} \alpha_k \phi_k(t) \ , \ \ t \in [0,1] \} \,, \tag{A2.1a}$$

$$H_N^{(r)} \triangleq \mathbb{R}^{n_s} \times L_N^{(r)} \,. \tag{A2.1b}$$

where $\alpha_k \in \mathbb{R}^m$, $\phi_k : [0,1] \to \mathbb{R}$ are the basis function $\phi_k(t) = B_{k,r,\mathbf{t}}(t)$, as defined in [4] and r is the order (one more than the degree) of the polynomials that make up the spline pieces. The subscript $\mathbf{t}$ is a knot sequence which we choose for our purposes to be $\mathbf{t} = \{ k/N \}_{k=-r+1}^{N+r-1}$ (note that, unlike in [4], our indexing does not start at k=1). With this knot sequence, the B-splines constitute a basis for the space of $r-2$ times continuously differentiable splines of order r that have breakpoints at times $t_k = k/N$, $k = \{0, \ldots, N\}$. Since splines are just piecewise polynomials between breakpoints with continuity and smoothness constraints at the breakpoints, $L_N^{(r)} \subset L_N^1$ and $H_N^{(r)} \subset H_N \triangleq \mathbb{R}^{n_s} \times L_N^1$, where L_N^1 is defined in Section 4 for representation **R1** with r-th order polynomial pieces. The control samples, $u[\tau_{k,j}]$, $k = 0, \ldots, N-1$, $j \in \mathbf{r}$, used by the RK integration method given in (4.3a,b) are related to the spline coefficients by $u[\tau_{k,j}] = \sum_{k=1}^{N+r-1} \alpha_k \phi_k(\tau_{k,j})$.

We will use B-splines normalized so that $\sum_{k=1}^{N+r-1} B_{k,r,\mathbf{t}}(t) = 1$ for all $t \in [0,1]$. These B-splines can be written in terms of the following recursion on the spline order r:

$$B_{k,r+1,\mathbf{t}}(t) = \frac{t - t_{k-r-1}}{t_{k-1} - t_{k-r-1}} B_{k,r,\mathbf{t}}(t+\Delta) + \frac{t_k - t}{t_k - t_{k-r}} B_{k+1,r,\mathbf{t}}(t+\Delta) \,, \ \ k = 1, \ldots, N+r \,, \ r \geq 1 \,, \tag{A2.2a}$$

where $\Delta = 1/N$ and

$$B_{k,1,t} = \begin{cases} 1\,, & t_{k-1} \le t \le t_k \\ 0\,, & \text{otherwise} \end{cases}\,, \quad k = 1, \dots, N\,. \tag{A2.2b}$$

For instance, cubic splines ($r = 4$) have the following basis functions [21]:

$$B_{k,4,t}(t) = \frac{1}{6\Delta^3} \begin{cases} (t - t_{k-4})^3\,, & t_{k-4} \le t \le t_{k-3} \\ \Delta^3 + 3\Delta^2(t - t_{k-3}) + 3\Delta(t - t_{k-3})^2 - 3(t - t_{k-3})^3\,, & t_{k-3} \le t \le t_{k-2} \\ 4\Delta^3 - 6\Delta(t - t_{k-2})^2 + 3(t - t_{k-2})^3\,, & t_{k-2} \le t \le t_{k-1} \\ \Delta^3 - 3\Delta^2(t - t_{k-1}) + 3\Delta(t - t_{k-1})^2 - (t - t_{k-1})^3\,, & t_{k-1} \le t \le t_k \end{cases}. \tag{A2.2c}$$

The domain of the B-splines extends outside of the range $t \in [0,1]$ for the purpose of construction only. The functions $u(t)$, given by (A2.1a), are defined only on $t \in [0,1]$. An important feature of B-splines is that the support of each individual basis function is only r intervals $[t_k, t_{k+1}]$. This is important for efficient computation of $u(t)$ from the spline coefficients and of the gradients of the cost and constraint functions.

We define the control constraint sets for the approximating problems as,

$$\mathbf{U}_N^{(r)} \triangleq \{\, u \in L_N^{(r)} \mid \alpha_k \in U\,,\ k = 1, \dots, N + r - 1 \,\} \tag{A2.3a}$$

$$\mathbf{H}_N^{(r)} \triangleq \mathbb{R}^{n_i} \times \mathbf{U}_N^{(r)}\,, \tag{A2.3b}$$

where, for this appendix, we assume that U, used to define $\mathbf{U}$ in (3.3a), is a cartesian product of the form

$$U \triangleq \overset{m}{\underset{i=1}{X}} [a_i, b_i]\,, \tag{A2.3c}$$

with $|a_i| < \infty$, $|b_i| < \infty$ and $a_i < b_i$. The approximating problems are thus:

$$\mathbf{CP}_N \qquad \min_{\eta \in \mathbf{H}_N^{(r)}} \{\, \psi_{o,N}(\eta) \mid \psi_{c,N}(\eta) \le 0 \,\}\,. \tag{A2.3d}$$

The functions $\psi_{o,N}(\eta)$ and $\psi_{c,N}(\eta)$ are defined as in (4.17). We will keep the definition of the optimality functions the same as given in Section 5. Note that the decision parameters for these problems transcribed into coefficient space, $\bar{H}_N^{(r)}$, are the coefficients α_k, $k = 1, \dots, N{+}r{-}1$ in the expansion of basis functions rather than the Nr control samples $u[\tau_{k,j}]$, $k = 0, \dots, N{-}1$, $j = 1, \dots, r$ for the approximating problems defined in Section 4. Thus, the number of decision parameters needed for splines is substantially less than the number needed for the same order general piecewise polynomials.

The next three results state properties of the spline subspaces that are needed to prove epiconvergence of the approximating problems to the original problem. Corollary A2.2. is a non-recursive restatement of the subdivision result presented in [33, Thm 3.1].

Proposition A2.1. (Nesting of Basis Functions) Given an integer $\rho \geq 1$, let $\mathbf{t} = \{ k / N \}_{k=1-\rho}^{k=N+\rho-1}$ and $\mathbf{t}' = \{ k / 2N \}_{k=1-\rho}^{k=2N+\rho-1}$. Then,

$$B_{k,\rho,\mathbf{t}}(t) = \frac{1}{2^{\rho-1}} \sum_{i=1}^{\rho+1} \sigma_{\rho,i} B_{2k-\rho+i-1,\rho,\mathbf{t}'}(t) \,, \quad t \in [0,1] \,, \quad k = 1, \ldots, N+\rho-1 \,, \tag{A2.4}$$

where $\sigma_{\rho,i}$ is the i-th coefficient of the polynomial $(t+1)^\rho$. We can take $B_{l,\rho,\mathbf{t}'}(t) \equiv 0$ if $l < 1-\rho$ or $l > 2N + \rho - 1$ since the proposition is stated only on $t \in [0,1]$.

Proof. We can prove (A2.4) by induction on ρ. It is clear from (A2.2b) that (A2.4) holds for $\rho = 1$. Now we will show that if (A2.4) holds for $\rho = r$, then it holds for $\rho = r+1$. From (A2.2a),

$$B_{k,r+1,\mathbf{t}}(t) = \frac{t - t_{k-r-1}}{r\Delta} B_{k,r,\mathbf{t}}(t + \Delta) + \frac{t_k - t}{r\Delta} B_{k+1,r,\mathbf{t}}(t + \Delta) \,, \quad k = 1, \ldots, N + r \,. \tag{A2.5a}$$

Substituting (A2.4) into this expression, letting $\Delta' = \Delta / 2$ and noting that $B_k(t + 2\Delta') = B_{k-1}(t + \Delta')$ and $t_k = t'_{2k}$, gives us

$$B_{k,r+1,\mathbf{t}}(t) = \frac{t - t'_{2(k-r-1)}}{2r\Delta'} \frac{1}{2^{r-1}} \sum_{i=1}^{r+1} \sigma_{r,i} B_{2k-r+i-1,r,\mathbf{t}'}(t + 2\Delta') + \frac{t'_{2k} - t}{2r\Delta'} \frac{1}{2^{r-1}} \sum_{i=1}^{r+1} \sigma_{r,i} B_{2k-r+i+1,r,\mathbf{t}'}(t + 2\Delta')$$

$$= \frac{1}{2^r} \left\{ \sigma_{r,1} \frac{t - t'_{2(k-r-1)}}{r\Delta'} B_{2k-r-1,r,\mathbf{t}'}(t + \Delta') + \sigma_{r,2} \frac{t - t'_{2(k-r-1)}}{r\Delta'} B_{2k-r,r,\mathbf{t}'}(t + \Delta') \right.$$

$$+ \sum_{j=2}^{r} \left[\sigma_{r,j+1} \frac{t - t'_{2(k-r-1)}}{r\Delta'} + \sigma_{r,j-1} \frac{t'_{2k} - t}{r\Delta'} \right] B_{2k-r+j-1,r,\mathbf{t}'}(t + \Delta')$$

$$\left. + \sigma_{r,r} \frac{t'_{2k} - t}{r\Delta'} B_{2k,r,\mathbf{t}'}(t + \Delta') + \sigma_{r,r+1} \frac{t'_{2k} - t}{r\Delta'} B_{2k+1,r,\mathbf{t}'}(t + \Delta') \right\}$$

$$= \frac{1}{2^r} \left\{ \sigma_{r,1} \frac{t - t'_{2(k-r-1)}}{r\Delta'} B_{2k-r-1} + \sigma_{r,1} \frac{t'_{2k-r-1} - t}{r\Delta'} B_{2k-r} + (\sigma_{r,1} + \sigma_{r,2}) \frac{t - t'_{2(k-r)-1}}{r\Delta'} B_{2k-r} \right.$$

$$+ \sum_{j=2}^{r} \left[(\sigma_{r,j-1} + \sigma_{r,j}) \frac{t'_{2(k-1)-r+j} - t}{r\Delta'} + (\sigma_{r,j} + \sigma_{r,j+1}) \frac{t - t'_{2(k-r-1)+j}}{r\Delta'} \right] B_{2k-r+j-1}$$

$$\left. + (\sigma_{r,r} + \sigma_{r,r+1}) \frac{t'_{2k-1} - t}{r\Delta'} B_{2k} + \sigma_{r,r+1} \frac{t - t'_{2k-r-1}}{r\Delta'} B_{2k} + \sigma_{r,r+1} \frac{t'_{2k} - t}{r\Delta'} B_{2k+1} \right\}$$

where we have abbreviated $B_{k,r,\mathbf{t}'}(t + \Delta')$ with B_k and we have used the following facts:

(i) since $r\sigma_{r,1} - \sigma_{r,2} = 0$,

$$\sigma_{r,2}(t - t'_{2(k-r-1)}) = (r\sigma_{r,1} - \sigma_{r,2})\Delta' + \sigma_{r,2}(t - t'_{2(k-r-1)})$$

$$= \sigma_{r,1}(t'_{2k-r-1} - t + t - t'_{2(k-r)-1}) + \sigma_{r,2}(t - t'_{2(k-r-1)} - \Delta')$$

$$= \sigma_{r,1}(t'_{2k-r-1} - t) + (\sigma_{r,1} + \sigma_{r,2})(t - t'_{2(k-r)-1}), \qquad \text{(A2.5b)}$$

(ii) since $\sigma_{r,r} - r\sigma_{r,r+1} = 0$,

$$\sigma_{r,r}(t'_{2k} - t) = \sigma_{r,r}(t'_{2k} - t) - (\sigma_{r,r} - r\sigma_{r,r+1})\Delta'$$

$$= \sigma_{r,r}(t'_{2k} - t - \Delta') + \sigma_{r,r+1}(t'_{2k-1} - t + t - t'_{2k-r-1})$$

$$= (\sigma_{r,r} + \sigma_{r,r+1})(t'_{2k-1} - t) + \sigma_{r,r+1}(t - t'_{2k-r-1}), \qquad \text{(A2.5c)}$$

(iii) and since $\sigma_{r,j} = \dfrac{d^{j-1}}{dt^{j-1}} \dfrac{(t+1)^r}{(j-1)!}\Bigg|_{t=0} = \dfrac{r(r-1)\cdots(r-j+2)}{(j-1)!}$, $j = 1, \ldots, r{+}1$ implies that $\sigma_{r,j-1}(r{+}2{-}j) - r\sigma_j + j\sigma_{j+1} = 0$, $j{=}2, \ldots, r$, we see that

$$\sum_{j=2}^{r} \sigma_{r,j-1}\frac{t'_{2k} - t}{r\Delta'} + \sigma_{r,j+1}\frac{t - t'_{2(k-r-1)}}{r\Delta'}$$

$$= \sum_{j=2}^{r} \sigma_{r,j-1}\frac{t'_{2k} - t}{r\Delta'} + \sigma_{r,j+1}\frac{t - t'_{2(k-r-1)}}{r\Delta'} - \frac{1}{r}(\sigma_{r,j-1}(r{+}2{-}j) - r\sigma_{r,j} + j\sigma_{r,j+1})$$

$$= \sum_{j=2}^{r} \sigma_{r,j-1}\frac{t'_{2k} - t}{r\Delta'} + \sigma_{r,j+1}\frac{t - t'_{2(k-r-1)}}{r\Delta'} - \sigma_{r,j-1}\frac{r{+}2{-}j}{r} - \sigma_{r,j+1}\frac{j}{r}$$

$$+ \sigma_{r,j}\frac{t'_{2(k-1)-r+j} - t + t - t'_{2(k-r-1)+j}}{r\Delta'}$$

$$= \sum_{j=2}^{r} \sigma_{r,j-1}\frac{t'_{2k} - t}{r\Delta'} - \sigma_{r,j-1}\frac{r{+}2{-}j}{r} - \sigma_{r,j}\frac{t'_{2(k-1)-r+j} - t}{r\Delta'} + \sigma_{r,j}\frac{t - t'_{2(k-r-1)+j}}{r\Delta'}$$

$$+ \sigma_{r,j+1}\frac{t - t'_{2(k-r-1)}}{r\Delta'} + \sigma_{r,j+1}\frac{j}{r}$$

$$= \sum_{j=2}^{r} (\sigma_{r,j-1} + \sigma_{r,j})\frac{t'_{2(k-1)-r+j} - t}{r\Delta'} + (\sigma_{r,j} + \sigma_{r,j+1})\frac{t - t'_{2(k-r-1)+j}}{r\Delta'} . \qquad \text{(A2.5d)}$$

Now, rearranging terms slightly, we get

$$B_{k,r+1,\mathbf{t}}(t) = \frac{1}{2^r}\left\{\sigma_{r,1}\frac{t - t'_{2(k-r-1)}}{r\Delta'}B_{2k-r-1} + \sigma_{r,1}\frac{t'_{2k-r-1} - t}{r\Delta'}B_{2k-r}\right.$$

$$+ \sum_{j=1}^{r}(\sigma_{r,j} + \sigma_{r,j+1})\left[\frac{t - t'_{2(k-r-1)+j}}{r\Delta'}B_{2k-r+j-1} + \frac{t'_{2k-r+j-1} - t}{r\Delta'}B_{2k-r+j}\right]$$

$$\left. + \sigma_{r,r+1}\frac{t - t'_{2k-r-1}}{r\Delta'}B_{2k} + \sigma_{r,r+1}\frac{t'_{2k} - t}{r\Delta'}B_{2k+1}\right\}. \tag{A2.5e}$$

Referring to (A2.2a) and noting that $\sigma_{r,1} = \sigma_{r+1,1} = 1$, $\sigma_{r,r+1} = \sigma_{r+1,r+2} = 1$, and $\sigma_{r,j} + \sigma_{r,j+1} = \sigma_{r+1,j+1}$, $j=1,\ldots,r+1$, we see that

$$B_{k,r+1,\mathbf{t}}(t) = \frac{1}{2^r}\sum_{i=1}^{r+2}\sigma_{r+1,i}B_{2k-r+i-2,r+1,\mathbf{t}'}(t), \tag{A2.5f}$$

which verifies that (A2.4) holds for $\rho = r+1$. □

Corollary A2.2. Let $r \geq 1$ and $\sigma_{r,i}$ be the i-th coefficient of the polynomial $(t+1)^r$. Then, given $u \in L_N^{(r)}$ with coefficients α_k, $k = 1,\ldots,N+r-1$, u is also a member of $L_{2N}^{(r)}$ with coefficients β_k, $k = 1,\ldots,2N+r-1$ given, for r odd, by

$$\beta_k = \frac{1}{2^{r-1}}\begin{cases}\displaystyle\sum_{i=1}^{\frac{r+1}{2}} \alpha_{(k+r)/2-i+1}\,\sigma_{r,2i-1}, & k \text{ odd} \\ \displaystyle\sum_{i=1}^{\frac{r+1}{2}} \alpha_{(k+r+1)/2-i}\,\sigma_{r,2i}, & k \text{ even}\end{cases}, \tag{A2.6a}$$

and, for r even, by

$$\beta_k = \frac{1}{2^{r-1}}\begin{cases}\displaystyle\sum_{i=1}^{\lceil\frac{r+1}{2}\rceil} \alpha_{(k+r)/2-i+1}\,\sigma_{r,2i-1}, & k \text{ even} \\ \displaystyle\sum_{i=1}^{\lfloor\frac{r+1}{2}\rfloor} \alpha_{(k+r-1)/2-i+1}\,\sigma_{r,2i}, & k \text{ odd}\end{cases}, \tag{A2.6b}$$

where $\lceil p \rceil$ is the smallest integer n such that $n \geq p$ and $\lfloor p \rfloor$ is the largest integer n such that $n \leq p$.

Proof. In the following, set $B_{k,r,\mathbf{t}'}(t) \equiv 0$ if $k < 1$ or $k > 2N+r-1$. From equation (A2.1a) and Proposition A2.1,

$$u(t) = \sum_{k=1}^{N+r-1} \alpha_k B_{k,r,\mathrm{t}'}(t) = \sum_{k=1}^{N+r-1} \alpha_k \frac{1}{2^{r-1}} \sum_{i=1}^{r+1} \sigma_{r,i} B_{2k-r+i-1,r,\mathrm{t}'}(t)$$

$$= \sum_{\substack{k'=1 \\ k' \text{ odd}}}^{2(N+r-1)} \alpha_{\frac{k'+1}{2}} \frac{1}{2^{r-1}} \sum_{i=1}^{r+1} \sigma_{r,i} B_{k'-r+i,r,\mathrm{t}'}(t)$$

$$= \sum_{\substack{k'=1 \\ k' \text{ odd}}}^{2(N+r-1)-1} \alpha_{\frac{k'+1}{2}} \frac{1}{2^{r-1}} \left(\sum_{j=1}^{\lceil \frac{r+1}{2} \rceil} \sigma_{r,2j-1} B_{k'-r+2j-1,r,\mathrm{t}'}(t) + \sum_{j=1}^{\lfloor \frac{r+1}{2} \rfloor} \sigma_{r,2j} B_{k'-r+2j,r,\mathrm{t}'}(t) \right)$$

$$= \sum_{k'=1}^{2(N+r-1)} \frac{1}{2^{r-1}} \begin{cases} \displaystyle\sum_{j=1}^{\lceil \frac{r+1}{2} \rceil} \alpha_{\frac{k'+1}{2}} \sigma_{r,2j-1} B_{k'-r+2j-1,r,\mathrm{t}'}(t) & k' \text{ odd} \\ \displaystyle\sum_{j=1}^{\lfloor \frac{r+1}{2} \rfloor} \alpha_{\frac{k'}{2}} \sigma_{r,2j} B_{k'-r+2j-1,r,\mathrm{t}'}(t) & k' \text{ even} \end{cases}$$

$$= \frac{1}{2^{r-1}} \begin{cases} \displaystyle\sum_{j=1}^{\lceil \frac{r+1}{2} \rceil} \sum_{k=2j-r}^{2N+r-3+2j} \alpha_{\frac{k+r}{2}+1-j} \sigma_{r,2j-1} B_{k',r,\mathrm{t}'}(t) & k+r \text{ even} \\ \displaystyle\sum_{j=1}^{\lfloor \frac{r+1}{2} \rfloor} \sum_{k=2j-r}^{2N+r-3+2j} \alpha_{\frac{k+r+1}{2}-j} \sigma_{r,2j} B_{k',r,\mathrm{t}'}(t) & k+r \text{ odd} \end{cases} \qquad \text{(A2.7a)}$$

Thus, if r is odd, we can write, abbreviating $B_{k,r,\mathrm{t}'}(t)$ with B_k,

$$u(t) = \frac{1}{2^{r-1}} \left\{ \sum_{k=2-r}^{2N+r-1} \begin{bmatrix} \alpha_{\frac{k+r}{2}} \sigma_{r,1} \\ \alpha_{\frac{k+r-1}{2}} \sigma_{r,2} \end{bmatrix} B_k + \sum_{k=4-r}^{2N+r+1} \begin{bmatrix} \alpha_{\frac{k+r}{2}-1} \sigma_{r,3} \\ \alpha_{\frac{k+r-1}{2}-1} \sigma_{r,4} \end{bmatrix} B_k + \cdots + \sum_{k=1}^{2(N+r-1)} \begin{bmatrix} \alpha_{\frac{k+1}{2}} \sigma_{r,r} \\ \alpha_{\frac{k}{2}} \sigma_{r,r+1} \end{bmatrix} B_k \right\},$$

where the top row is for k odd and the bottom row is for k even. If r is even, we can write

$$u(t) = \frac{1}{2^{r-1}} \left\{ \sum_{k=2-r}^{2N+r-1} \begin{bmatrix} \alpha_{\frac{k+r}{2}} \sigma_{r,1} \\ \alpha_{\frac{k+r-1}{2}} \sigma_{r,2} \end{bmatrix} B_k + \sum_{k=4-r}^{2N+r+1} \begin{bmatrix} \alpha_{\frac{k+r}{2}-1} \sigma_{r,3} \\ \alpha_{\frac{k+r-1}{2}-1} \sigma_{r,4} \end{bmatrix} B_k + \cdots + \sum_{k=1}^{2(N+r)-1} \begin{bmatrix} \alpha_{\frac{k}{2}} \sigma_{r,r+1} \\ 0 \end{bmatrix} B_k \right\},$$

where, the top row is for k even and the bottom row is for k odd. Now, by collecting the terms for $k \in \{ 1, \ldots, 2N+r-1 \}$ and forming the expression

$$u(t) = \sum_{k=1}^{2N+r-1} \beta_k B_{k,r,t'}(t) , \qquad \text{(A2.7b)}$$

we see that the coefficients β_k are as given by (A2.6a,b). □

Lemma A2.3. Let $\mathbf{N} = \{ 2^n \}_{n=1}^{\infty}$. Then, $L_{N_1}^{(r)} \subset L_{N_2}^{(r)}$ for any $N_1, N_2 \in \mathbf{N}$ such that $N_1 < N_2$. Furthermore,

(a) Given $\eta = (\xi, u) \in \mathbf{H}$ and $N = 2^n < \infty$, there exists $j_n \in \mathbf{N}$, $j_n < \infty$ and $\eta_{j_n} \in \mathbf{H}_{j_n}^{(r)}$ such that $\| \eta - \eta_{j_n} \| < 1/N$.

(b) Suppose there is a sequence $\{ u_N \}_{N \in \mathbf{N}}$ such that $u_N \in \mathbf{U}_N^{(r)}$ and $u_N \to u$. Then $u \in \mathbf{U}$.

Proof. The nesting of the subspaces follows directly from Corollary A2.2.

(a) From (A2.3c), U is given by a cartesian product of the form $\mathop{\mathsf{X}}_{i=1}^{m} [a_i, b_i]$, with $|a_i| < \infty$, $|b_i| < \infty$ and $a_i < b_i$. Since $u \in L_2^m[0,1]$, there exists $u'_\varepsilon \in C^m[0,1]$ such that $\| u - u'_\varepsilon \|_2 \le \delta = 2/(5+m)N$, [27, Theorem 3.14, p. 69]. Now, with ${}^i u_\varepsilon$ denoting the i-th component of u_ε, define the function u_ε as follows:

$${}^i u_\varepsilon(t) = \begin{cases} b_i - \delta & \text{if } \; {}^i u'_\varepsilon(t) > b_i - \delta , \\ {}^i u'_\varepsilon(t) & \text{if } \; a_i + \delta \le {}^i u'_\varepsilon(t) \le b_i - \delta , \quad i \in \mathbf{m} , \; t \in [0,1] , \\ a_i + \delta & \text{if } \; a_i + \delta < {}^i u'_\varepsilon(t) , \end{cases} \qquad \text{(A2.8a)}$$

Note that, because $a_i \le {}^i u(t) \le b_i$ for all $i \in \mathbf{m}$, $t \in [0,1]$,

$$\| u - u_\varepsilon \|_2^2 = \int_0^1 \sum_{i=1}^{m} ({}^i u(t) - {}^i u_\varepsilon(t))^2 dt \le \int_0^1 \sum_{i=1}^{m} (({}^i u(t) - {}^i u'_\varepsilon(t))^2 + \delta^2) \, dt = \| u - u'_\varepsilon \|_2^2 + m\delta^2 . \qquad \text{(A2.8b)}$$

Thus, $\| u - u_\varepsilon \|_2 \le (1+m)\delta$. Now, $u_\varepsilon(\cdot)$ is a continuous function on a compact interval, hence uniformly continuous. This implies that, for each $i \in \mathbf{m}$, the modulus of continuity for ${}^i u_\varepsilon$, $\omega({}^i u_\varepsilon, \sigma) \triangleq \max \{ | {}^i u_\varepsilon(t_1) - {}^i u_\varepsilon(t_2) | \; | \; |t_1 - t_2| \le \sigma \}$, goes to zero as $\sigma \to 0$. Thus, by [4, Theorem XII.1, p. 170], there exists an integer $N_1 = 2^{n_1} < \infty$ and $u_{N_1} \in L_{N_1}^{(\rho)}$ such that

$$\| u_\varepsilon - u_{N_1} \|_2 \le \| u_\varepsilon - u_{N_1} \|_\infty \le \frac{\delta}{2} = \frac{1}{(5} + m)N . \qquad \text{(A2.8c)}$$

Since u_{N_1} is also uniformly continuous, there exists $n_2 \in \mathbf{N}$, $n_1 < n_2 < \infty$, such that, with $N_2 = 2^{n_2}$,

$$\| u_{N_1}(t_1) - u_{N_1}(t_2) \| \le \frac{\delta}{D_{r,\infty} - 1} , \quad \forall t \in [0,1] \text{ such that } |t_1 - t_2| \le (\rho - 1)/N_2 , \qquad \text{(A2.8d)}$$

where $1 \le D_{r,\infty} < \infty$ is as given on page 155 of [4]. Now, for $k = 1, \ldots, N_2 + \rho$, define the intervals $T_k \triangleq [t_{k-\rho}, t_{k-1}]$, with $t_k = k/N_2$, and define the quantities ${}^i M_k = \max_{t \in T_k} {}^i u_{N_1}(t)$, and

${}^i m_k = \min_{t \in T_k} {}^i u_{N_1}(t)$. Since for $t_1, t_2 \in T_k$, $|t_1 - t_2| \leq (\rho - 1)/N_2$, we see that

$$ {}^i M_k - {}^i m_k = \max_{t_1, t_2 \in T_k} \| {}^i u_{N_1}(t_1) - {}^i u_{N_1}(t_2) \| \leq \frac{\delta}{D_{r,\infty} - 1} \ , \ \text{for } t_1, t_2 \in T_k \ , \ i \in \mathbf{m} \, . \qquad \text{(A2.8e)} $$

Thus,

$$ D_{r,\infty} \leq \frac{\delta}{{}^i M_k - {}^i m_k} + 1 \ , \ \forall i \in \mathbf{m} \, . \qquad \text{(A2.8f)} $$

Next, since $L_{N_1} \subset L_{N_2}$ by Corollary A2.2, $u_{N_1} \in L_{N_2}$. Hence, there exists $\{ \alpha_k \}_{k=1}^{N_2+\rho} \subset \mathbb{R}^m$ such that $u_{N_1}(t) = \sum_{k=1}^{N_2+\rho} \alpha_k \phi_k(t)$. Thus, by [4, Corollary XI.2, p. 156],

$$ | {}^i \alpha_k - \frac{{}^i M_k + {}^i m_k}{2} | \leq D_{r,\infty} \frac{{}^i M_k - {}^i m_k}{2} \leq \tfrac{1}{2}\delta + \frac{{}^i M_k - {}^i m_k}{2} \ , \qquad \text{(A2.8g)} $$

where we used (A2.8f) for the second inequality. Therefore, $-\tfrac{1}{2}\delta + {}^i m_k \leq {}^i \alpha_k \leq \tfrac{1}{2}\delta + {}^i M_k$. But, from (A2.8c), we see that ${}^i M_k \triangleq \max_{t \in T_k} {}^i u_{N_1}(t) \leq \max_{t \in T_k} {}^i u_\varepsilon(t) + \tfrac{1}{2}\delta = b_i - \tfrac{1}{2}\delta$ and ${}^i m_k \triangleq \min_{t \in T_k} {}^i u_{N_1}(t) \geq \min_{t \in T_k} {}^i u_\varepsilon(t) - \tfrac{1}{2}\delta = a_i + \tfrac{1}{2}\delta$. Thus, $a_i \leq {}^i \alpha_k \leq b_i$ which implies that $\alpha_k \in U$. Finally, by (A2.8b) and (A2.8c),

$$ \| u - u_{N_1} \|_2 \leq \| u - u'_\varepsilon \|_2 + \| u'_\varepsilon - u_\varepsilon \|_2 + \| u_\varepsilon - u_{N_1} \|_2 \leq \delta + (1 + m)\delta + \frac{\delta}{2} = \frac{1}{N} \ , \qquad \text{(A2.8h)} $$

since $\delta = 2/(5+m)N$. Thus, the proposition holds with $j_n = 2^{n_2}$ and $\eta_{j_n} = (\xi, u_{N_1}) \in H_{j_n}^{(\rho)}$.

(b) Referring to [4, Corollary XI.1, p. 155], we see that $\mathbf{U}_N^{(r)} \subset \mathbf{U}$ and $\mathbf{U}$ is closed. □

Remark A2.4. We see from Lemma A2.3*(b)* and the definition of $\mathbf{H}_N$ in Section 4 for representation **R1** that $\mathbf{H}_N^{(r)} \subset \mathbf{H} \cap H_N^{(r)} \subset \mathbf{H} \cap H_N \subset \mathbf{H}_N$. Hence, control constraint violations are possible for $\eta \in \mathbf{H}_N$ but not for $\eta \in \mathbf{H}_N^{(r)}$.

Theorem A2.5. (Epiconvergence) Suppose that Assumptions 3.1*(a)*, 4.1 and 4.9 hold. Let $\mathbf{N} = \{ 2^n \}_{n=1}^{\infty}$. Then, the problems $\{ \mathbf{CP}_N \}_{N \in \mathbf{N}}$ converge epigraphically to the problem $\mathbf{CP}$ as $N \to \infty$.

Proof. Given $\eta \in \mathbf{H}$, there exists, by Assumption 4.9, a sequence $\{ \eta_N \}_{N=1}^{\infty}$ such that $\eta_N \in \mathbf{H}$, $\eta_N \to \eta$ and $\psi_c(\eta_N) < 0$. By Lemma A2.3*(a)*, for each $N = 2^n$, there exists $j_n \in \mathbf{N}$ and $\eta_{j_n}' \in \mathbf{H}_{j_n}^{(r)}$ such that $\| \eta_N - \eta_{j_n}' \| \leq 1/N$. It now follows from the proof in Theorem 4.10 that part *(a)* of Definition 2.1 is satisfied. That part *(b)* of Definition 2.1 is satisfied follows from Lemma A2.3*(b)* and the proof in Theorem 4.10. □

To show consistency of approximations, what remains is to compute the gradients of the cost and constraint functions with respect to elements of $L_N^{(r)}$ and show that the optimality functions for the approximating problems hypoconverge to the optimality function for the original problem. To compute the gradient $\nabla_u f_N^\nu(\eta)$, $\nu \in \mathbf{q}$, we first define the space $\bar{L}_N^{(r)} \triangleq \mathbb{R}^{N+r-1} \times \mathbb{R}^m$ and the map

$$S_{N,r} : L_N^{(r)} \rightarrow \bar{L}_N^{(r)} , \tag{A2.9}$$

which takes elements $u = \sum_{k=1}^{N+r-1} \alpha_k \phi_k(t)$ and maps them to $\bar{\alpha} = \{ \alpha_k \}_{k=1}^{N+r-1}$, with $\alpha_k \in \mathbb{R}^m$. We also define $\bar{H}_N^{(r)} \triangleq \mathbb{R}^{n_x} \times L_N^{(r)}$. It is clear that $S_{N,r}$ is a linear bijection. Proceeding as in Section 4, we define the inner product on $\bar{L}_N^{(r)}$ in the following way. Given $\bar{\alpha}, \bar{\beta} \in \bar{L}_N^{(r)}$, let $u = S_{N,r}^{-1}(\bar{\alpha})$ and $v = S_{N,r}^{-1}(\bar{\beta})$. The inner product must satisfy

$$\langle \bar{\alpha}, \bar{\beta} \rangle_{\bar{L}_N^{(r)}} = \langle u , v \rangle_{L_2} = \int_0^1 \langle \sum_{k=1}^{N+r-1} \alpha_k \phi_k(t) , \sum_{l=1}^{N+r-1} \beta_l \phi_l(t) \rangle dt$$

$$= \sum_{k=1}^{N+r-1} \sum_{l=1}^{N+r-1} \langle \alpha_k , \beta_l \rangle \int_0^1 \phi_k(t) \phi_l(t) \, dt = \langle \bar{\alpha} \mathbf{M}_\alpha, \bar{\beta} \rangle_{l_2} , \tag{A2.10a}$$

where the inner product on the right hand side is the standard inner product on $\mathbb{R}^{N+r-1} \times \mathbb{R}^m$. Thus, $\mathbf{M}_\alpha$ is the $(N + r - 1) \times (N + r - 1)$ matrix whose k, l-th entry is given by

$$[\mathbf{M}_\alpha]_{k,l} = \int_0^1 \phi_k(t) \phi_l(t) \, dt \; . \tag{A2.10b}$$

An alternate means of determining $\mathbf{M}_\alpha$ is to make use of the fact that $L_N^{(r)} \subset L_N^1$. This will allow us to use the results for L_N^1 in Section 5 to show consistency. Let $\mathbf{M}_N$ be as defined in (4.9b) with $M = M_1$, the quadrature matrix for representation **R1**. Recall from Section 4 that, given $u \in L_N^1$, $V_{\mathbf{A},N}(u) = \bar{u} \in \bar{L}_N^1$. Thus, from (A2.1a), the composite map $V_{\mathbf{A},N} \circ S_{N,r}^{-1}(\bar{\alpha}) = \bar{\alpha} \Phi_{\mathbf{A},N}^T$ where the $(kr+j, l)$-th entry of the $Nr \times (N + r - 1)$ matrix $\Phi_{\mathbf{A},N}$ is $\phi_l(\tau_{k,j})$, $k = 0, \ldots, N-1$, $j = 1, \ldots, r$, and $l = 1, \ldots, N + r - 1$. Thus,

$$\langle u , v \rangle_{L_2} = \langle S_{N,r}(u) \mathbf{M}_\alpha, S_{N,r}(v) \rangle_{l_2} = \langle V_{\mathbf{A},N}(u) \mathbf{M}_N , V_{\mathbf{A},N}(v) \rangle_{l_2}$$

$$= \langle V_{\mathbf{A},N} \circ S_{N,r}^{-1} \circ S_{N,r}(u) \mathbf{M}_N , V_{\mathbf{A},N} \circ S_{N,r}^{-1} \circ S_{N,r}(v) \rangle_{l_2}$$

$$= \langle S_{N,r}(u) \Phi_{\mathbf{A},N}^T \mathbf{M}_N , S_{N,r}(v) \Phi_{\mathbf{A},N}^T \rangle_{l_2} . \tag{A2.11a}$$

Therefore,

$$\mathbf{M}_\alpha = \Phi_{\mathbf{A},N}^T \mathbf{M}_N \Phi_{\mathbf{A},N} \; . \tag{A2.11b}$$

It is not obvious that (A2.11b) is equivalent to (A2.10b), and hence, independent of the Butcher array **A**. To see that this is so, notice that the k, j-th element of $\mathbf{M}_\alpha$, as given in (A2.11b), is

$$[\mathbf{M}_\alpha]_{k,l} = \left[\; \phi_k(\tau_{0,1}) \cdots \phi_k(\tau_{0,r}) \cdots \phi_k(\tau_{N-1,1}) \cdots \phi_k(\tau_{N-1,r}) \right] \mathbf{M}_N \begin{bmatrix} \phi_l(\tau_{0,1}) \\ \vdots \\ \phi_l(\tau_{N-1,r}) \end{bmatrix} . \tag{A2.11c}$$

This is just the inner-product of $\phi_k(t)$ and $\phi_l(t)$ in L_N^1. Hence, because of the way M_1 was defined in

Section 4, $[\mathbf{M}_\alpha]_{k,l} = \langle \phi_k(t), \phi_l(t) \rangle_{L_2}$.

We are now in a position to compute the gradients of the cost and constraint functions with respect to elements of $L_N^{(r)}$. From Theorem 5.1 we have, for $\nu \in \mathbf{q}$, $\eta = (\xi, u) \in H_N^{(r)}$, and $\delta u \in H_N^{(r)}$,

$$\langle S_{N,r}(\nabla_u f_N^\nu(\eta))\mathbf{M}_\alpha, S_{N,r}(\delta u) \rangle_{l_2} = \langle \bar{\gamma}_\alpha^\nu(\bar{\alpha}), \delta\bar{\alpha} \rangle_{l_2}, \tag{A2.12a}$$

where $\delta\bar{\alpha} = S_{N,r}(\delta u)$, $\bar{\alpha} = S_{N,r}(\eta)$ and

$$\bar{\gamma}_\alpha^\nu(\bar{\alpha}) = \bar{\gamma}_u^\nu(\bar{\eta})\,\Phi_{\mathbf{A},N}\,, \tag{A2.12b}$$

with $\bar{\eta} = W_{\mathbf{A},N}(S_{N,r}^{-1}(\bar{\alpha}))$ and $\bar{\gamma}_u^\nu(\bar{\eta})$ defined by (5.5c). Thus, for $\eta \in L_N^{(r)}$ and $\bar{\alpha} = S_{N,r}(\eta)$,

$$\nabla_u f_N^\nu(\eta) = S_{N,r}^{-1}\left(\bar{\gamma}_\alpha^\nu(\bar{\alpha})\,\mathbf{M}_\alpha^{-1}\right). \tag{A2.13a}$$

Since $S_{N,r}^{-1}(\bar{\alpha})$ is a spline for any $\bar{\alpha} \in \bar{L}_N^{(r)}$, (A2.13a) shows that $\nabla_u f_N^\nu(\eta)$, $\nu \in \mathbf{q}$, is a spline for all $\eta \in H_N^{(r)}$. Note that, from (A2.11b) and (A2.13a) that, since $V_{\mathbf{A},N} \circ S_{N,r}^{-1}(\bar{\alpha}) = \bar{\alpha}\Phi_{\mathbf{A},N}$,

$$V_{\mathbf{A},N}(\nabla_u f_N^\nu(\eta)) = \bar{\gamma}_u^\nu(\bar{\eta})\Phi_{\mathbf{A},N}\,[\Phi_{\mathbf{A},N}^T\,\mathbf{M}_N\,\Phi_{\mathbf{A},N}]^{-1}\Phi_{\mathbf{A},N}^T = \bar{\gamma}_u^\nu(\bar{\eta})\,\mathbf{M}_N^{-1}\,, \tag{A2.13b}$$

since $\Phi_{\mathbf{A},N}$ has full rank. Equation (A2.13b) is the expression given in Theorem 5.1.

To show convergence of the gradients, we note that $L_N^{(r)} \subset L_N^1$. Therefore, by Theorem 5.6, there exists $\kappa < \infty$ such that $\|\nabla_u f_N^\nu(\eta) - \nabla_u f^\nu(\eta)\| \le \kappa/N$ for all $\eta \in H_N^{(r)} \subset H_N$. Therefore, the optimality functions hypoconverge by the result of Theorem 5.9. This, along with Theorem A2.5, shows that the approximating problems $\mathbf{CP}_N$, with feasible sets $\mathbf{H}_N^{(r)}$ and optimality functions θ_N given by (5.8a) using (A2.13a) as the expression for the gradients, are consistent approximations to $(\mathbf{CP}, \theta)$. We state this result as a theorem:

Theorem A2.6. Suppose that Assumptions 3.1, 4.1 and 4.9 and equation (5.18) hold. Let $\mathbf{N} = \{ 2^n \}_{n=1}^{\infty}$. Then, with $\mathbf{CP}_N$ as defined in (A2.3d) and θ_N as defined in (5.8a), the family of approximating pairs $(\mathbf{CP}_N, \theta_N)$, $N \in \mathbf{N}$, constitute consistent approximations for the pair $(\mathbf{CP}, \theta)$. □

Example (Linear Splines --- hat functions)

In this case, $r = 2$ and the basis functions are given by

$$\phi_k(t) = \begin{cases} (t - t_{k-2})/\Delta & \text{if } t \in [t_{k-2}, t_{k-1}] \\ (t_k - t)/\Delta & \text{if } t \in [t_{k-1}, t_k] \end{cases} \quad . \tag{A2.14}$$

Let $u, v \in L_N^{(r)}$ and $\bar{\alpha} = S_{N,r}(u)$ and $\bar{\beta} = S_{N,r}(v)$. Since these hat functions have a support of only two time intervals (2Δ), $\mathbf{M}_\alpha$, given by equation (A2.10b), is

$$\mathbf{M}_\alpha = \frac{\Delta}{6} \begin{bmatrix} 2 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & & \ddots & & \\ & & & & 4 & 1 \\ & & & & 1 & 2 \end{bmatrix} \quad . \tag{A2.15}$$

Example (Cubic Splines)

In this case, r=4 and the basis functions are given by (A2.2c). Assuming $k \le l$, $\int_0^1 \phi_k(t)\phi_l(t)\,dt = \int_a^b \phi_k(t)\phi_l(t)\,dt$ where $a = \max\{0, t_{l-4}\}$ and $b = \min\{t_k, 1\}$ since each B-spline has support of width 4Δ. In particular, $[\mathbf{M}_\alpha]_{k,l} = 0$ if $|k-l| > 3$. Thus, from (A2.10b),

$$\mathbf{M}_\alpha = \frac{\Delta}{5040} \begin{bmatrix} 20 & 129 & 60 & 1 & & & & \\ 129 & 1208 & 1062 & 120 & 1 & & & \\ 60 & 1062 & 2396 & 1191 & 120 & 1 & & \\ 1 & 120 & 1191 & 2416 & 1191 & 120 & 1 & \\ & 1 & 120 & 1191 & 2416 & 1191 & 120 & 1 \\ & & \ddots & & \ddots & & & \ddots \end{bmatrix} \quad . \tag{A2.16}$$

Note that, for $r \ge 2$, $\mathbf{M}_\alpha^{-1}$ is a dense matrix. But, for $\eta = (\xi, u) \in H_N^{(r)}$, we can find $\bar{d}^\nu(\eta) \triangleq S_{N,r}(\nabla_u f_N^\nu(\eta))$ efficiently by solving

$$\bar{d}^\nu(\eta)\mathbf{M}_\alpha = \bar{\gamma}_\alpha^\nu(\bar{\alpha}) \, , \tag{A2.17}$$

where $\bar{\alpha} = S_{N,r}(u)$. We can efficiently solve (A2.17) using the Cholesky decomposition of $\mathbf{M}_\alpha$ which can be computed off-line.

REFERENCES

[1] U. Ascher and G. Bader, *Stability of collocation at gaussian points*, SIAM J. Numer. Anal., 23 (1986), pp. 412-422.

[2] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman, London, 1984.

[3] T. E. Baker and E. Polak, *On the optimal control of systems described by evolution equations*, SIAM J. Control and Optimization, 32 (1994), pp. 224-260.

[4] Carl de Boor, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.

[5] Haim Brezis, *Analyse Fonctionnelle*, Masson, Paris, 1983.

[6] B. M. Budak, E. M. Berkovich and E. N. Solov'eva, *Difference approximations in optimal control problems*, SIAM J. Control, 7 (1969), pp. 18-31.

[7] J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, John Wiley and Sons, England, 1987.

[8] P. G. Ciarlet, M. H. Schultz, and R. S. Varga, *Numerical methods of high-order accuracy for nonlinear boundary value problems*, Numerische Mathematik, 9 (1967) pp. 394-430.

[9] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[10] J. Cullum, *Discrete approximations to continuous optimal control problems*, SIAM J. Control, 7 (1969), pp. 32-49.

[11] J. Cullum, *An Explicit procedure for discretizing continuous, optimal control problems*, Journal of Optimization Theory and Applications, 8 (1971) pp. 15-35.

[12] J. E. Cuthrell, L. T. Biegler, *On the optimization of differential-algebraic process systems*, AIChE Journal, 33 (1987), pp. 1257-1270.

[13] James, W. Daniel, *The Approximate Minimization of Functionals*, Prentice-Hall, New Jersey, 1971.

[14] S. Dolecki, G. Salinetti and R.J.-B. Wets, *Convergence of functions: equisemicontinuity*, Transactions of the American Mathematical Society 276 (1983), 409-429.

[15] J. C. Dunn, *Diagonally modified conditional gradient methods for input constrained optimal control problems*, SIAM J. Control Optim. 24 (1986), pp. 1177-1191.

[16] William W. Hager, *Rates of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal, 13 (1976), pp. 449-472.

[17] L. He and E. Polak, *An Optimal diagonalization strategy for the solution of a class of optimal design problems*, IEEE, Trans. on Autom. Contr., 35 (1990), pp. 258-267.

[18] J. D. Lambert, *Numerical Methods for Ordinary Differential Systems*, John Wiley and Sons, England, 1991.

[19] B. Sh. Mordukhovich, *On Difference approximations of optimal control systems*, J. Appl. Math. Mech, 42 (1978), pp. 452-461.

[20] B. Sh. Mordukhovich, *Methods of Approximation in Optimal Control Problems*, (in Russian) Nauka, Moscow, 1988.

[21] C. P. Neuman and A. Sen, *A Suboptimal control algorithm for constrained problems using cubic splines*, Automatica, 9 (1973), pp. 601-613.

[22] E. Polak and L. He, *Unified steerable phase I-phase II method of feasible directions for semi-infinite optimization*, Journal of Optimization Theory and Applications, 69 (1991), pp. 83-107.

[23] E. Polak, *Computational Methods in Optimization*, Academic Press, 1971.

[24] E. Polak, *On the use of consistent approximations in the Solution of semi-infinite optimization and optimal control problems*, Math. Prog., 62 (1993), pp. 385-415.

[25] E. Polak and L. He, *Rate-preserving discretization strategies for semi-infinite programming and optimal control*, SIAM J. Control and Optimization, 30 (1992), pp. 548-572.

[26] G. W. Reddien, *Collocation at gauss points as a discretization in optimal control*, SIAM J. Control and Optimization, 17 (1979), pp. 298-306.

[27] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, 1987.

[28] R. Schere and H Turke, *Reflected and transposed runge-kutta methods*, BIT 23 (1983), pp. 262-266.

[29] O. Stryk and R. Bulirsch, *Direct and indirect methods for trajectory optimization*, Annals of Operations Research, 37 (1992), pp. 357-373.

[30] L. J. Williamson and E. Polak, *Relaxed controls and the convergence of optimal control algorithms*, SIAM J. Control, 14 (1976), pp. 737-757.

[31] V. Veliov, *Second-order discrete approximations to linear differential inclusions*, SIAM J. Numer. Anal., 29 (1992), pp. 439-451.

[32] J. Vlassenbroeck and R. V. Dooren, *A Chebyshev technique for solving nonlinear optimal control problems*, IEEE Trans. Autom. Cntrl., 33 (1988), pp. 333-340.

[33] J. M. Lane and R. F. Riesenfeld, *A theoretical development for the computer generation of piecewise polynomial sufraces*, IEEE Trans. Patern Anal. and Machine Intelligence, 2 (1980), pp. 35-46.