2002-08-05

A Perspective on Methods for Trajectory Optimization

Ross, I. Michael; Fahroo, Fariba

The American Institute of Aeronautics and Astronautics (AIAA)

Ross, I. Michael, and Fariba Fahroo. "A perspective on methods for trajectory optimization, aiaas." AIAA Guidance, Navigation, and Control Conference and Exhibit, 5-8 August 2002, Monterey, California.. The American Institute of Aeronautics and Astronautics (AIAA), 2002.

https://hdl.handle.net/10945/29671

AIAA/AAS Astrodynamics Specialist Conference and Exhibit
5-8 August 2002, Monterey, California

AIAA 2002-4727

A PERSPECTIVE ON METHODS FOR TRAJECTORY OPTIMIZATION

I. Michael Ross * and Fariba Fahroo †
Naval Postgraduate School, Monterey, CA 93943

In recent years there has been an upsurge of purported new methods for trajectory generation and optimization each promising many advantages over another. Frequently, one has to deal with new, and sometimes confusing, terminologies such as inverse methods, differential inclusions, differential flatness, collocation, pseudospectral methods, higher-order methods and so on. In this paper we provide a mathematical framework that distinguishes or blends the various approaches. This framework is facilitated by distinguishing a transformation from the discretization rather than bundling them together. Two clear layers emerge with two types of convergence notions. A Covector Mapping Principle is enunciated which facilitates the definition of a Complete Method. Many competing claims and issues can now be resolved. A surprising conclusion that can be drawn from this is that only a few fundamental methods for trajectory generation and optimization of complex dynamical systems have been studied in the literature. Hence many basic research questions remain unanswered.

1 Introduction

Betts[1,2] provides an excellent review and survey of the many numerical methods for trajectory optimization. The purpose of this paper is not to provide a survey, but to provide a perspective on the various methods for trajectory optimization. A vast array of sometimes confusing terminologies abound in the literature: *inverse methods, dynamic inversion, differential inclusions, differential flatness and so on*. While it may be clear that these are valuable ***concepts***, an important question is: do these concepts translate to new ***numerical methods***? Can a framework be put forth that either distinguishes a method or unifies it with some other? The purpose of this exercise is not to promote or disparage a method but to identify how it fits within a mathematical framework. Apparently, Betts[2] foresaw this as he writes, "... one may expect many of the best features of seemingly disparate techniques to merge, forming still more powerful methods."

In order to facilitate the discussion of various methods, we articulate the following *basic* problem:

Problem B

Let $\mathbf{x} \in \mathbb{R}^{N_x}$ and $\mathbf{u} \in \mathbb{R}^{N_u}$. Determine the state-control function-pair, $\{\mathbf{x}(\cdot), \mathbf{u}(\cdot)\}$ that minimize the fixed-time Lagrange cost functional,

$$J[\mathbf{x}(\cdot), \mathbf{u}(\cdot)] = \int_{\tau_0}^{\tau_f} F(\mathbf{x}(\tau), \mathbf{u}(\tau), \tau) d\tau \tag{1}$$

subject to the dynamic constraints,

$$\dot{\mathbf{x}}(\tau) = \mathbf{f}(\mathbf{x}(\tau), \mathbf{u}(\tau), \tau) \tag{2}$$

and end-point constraints,

$$\mathbf{x}(\tau_0) = \mathbf{x}_0 \tag{3}$$

$$\mathbf{e}\big(\mathbf{x}(\tau_f)\big) = \mathbf{0} \tag{4}$$

where it is assumed that the functions,

$$F: \quad \mathbb{R}^{N_x} \times \mathbb{R}^{N_u} \times \mathbb{R} \to \mathbb{R} \tag{5}$$

$$\mathbf{f}: \quad \mathbb{R}^{N_x} \times \mathbb{R}^{N_u} \times \mathbb{R} \to \mathbb{R}^{N_x} \tag{6}$$

$$\mathbf{e}: \quad \mathbb{R}^{N_x} \to \mathbb{R}^{N_e} \tag{7}$$

are continuously differentiable with respect to their arguments. The simplicity of the problem formulation posed above begs an important question: If a method is capable of solving this problem, can it be generalized to solving vastly more complex problems that may be abstracted as,

Problem G

Determine the state-control function-pair, $\{\mathbf{x}(\cdot), \mathbf{u}(\cdot)\}$, design parameters $\mathbf{p}$, and the clock times τ_0 and τ_f that minimize the Bolza cost functional,

$$J[\mathbf{x}(\cdot), \mathbf{u}(\cdot), \tau_0, \tau_f; \mathbf{p}] = E(\mathbf{x}(\tau_0), \mathbf{x}(\tau_f), \tau_0, \tau_f; \mathbf{p}) + \int_{\tau_0}^{\tau_f} F(\mathbf{x}(\tau), \mathbf{u}(\tau), \tau; \mathbf{p}) d\tau \tag{8}$$

subject to the constraints,

$$\dot{\mathbf{x}} \in \mathcal{F}(\mathbf{x}, \mathbf{u}, \tau; \mathbf{p}) \tag{9}$$

$$\mathbf{x} \in \mathcal{X}, \quad \mathbf{u} \in \mathcal{U}, \quad \mathbf{p} \in \mathcal{P}, \quad \mathcal{X} \times \mathcal{U} \times \mathcal{P} \in \mathcal{H} \tag{10}$$

where all functions are piecewise differentiable, the denoted sets may be given explicitly by equalities and inequalities and the number of discontinuities in the state and control variables are finite. Although this *general* problem is stated abstractly, it is arguably the most important problem from a practical perspective of solving "highly complex problems" arising

*Associate Professor, Department of Aeronautics and Astronautics. Associate Fellow, AIAA. E-mail: imross@nps.navy.mil
†Associate Professor, Department of Mathematics. Senior Member, AIAA. E-mail: ffahroo@nps.navy.mil

in aeronautics, astronautics, robotics and many other disciplines where "nonsmoothness" and nonlinearities abound. See Bryson[3] for a wide range of solved problems.

Our discussion will be focused on the capability of methods that can solve the range of problems implied by the simplicity in the formulation of B to the complexity of G.

2 Transformations

We begin with Problem B as it is the most familiar. The necessary conditions are given by the famous Minimum Principle due to Pontryagin and others.[4] In terms of the control Hamiltonian defined as

$$H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u}, \tau) = F + \boldsymbol{\lambda}^T \mathbf{f} \tag{11}$$

the necessary optimality conditions are,

$$\frac{\partial H}{\partial \mathbf{u}} = \mathbf{0}, \tag{12}$$

where $\boldsymbol{\lambda}(t)$ is the Lagrange multiplier governed by the adjoint equation (costate dynamics) and the transversality conditions

$$\dot{\boldsymbol{\lambda}} = -\frac{\partial H}{\partial \mathbf{x}} \tag{13}$$

$$\boldsymbol{\lambda}(\tau_f) = \left(\frac{\partial \mathbf{e}}{\partial \mathbf{x}(\tau_f)}\right)^T \boldsymbol{\nu}_f \tag{14}$$

These necessary conditions ***do not solve the problem*** (at least from a numerical perspective); rather they ***transform*** it to another problem in a higher dimensional space, viz., the primal-dual space. This transformation (known as the Legendre-Fenchel transformation[5]) is achieved by way of the Hamiltonian and is said to dualize the problem. This transformed problem is the two-point boundary-value-problem (BVP):

Problem B^λ

Determine the state-control-costate function-triple, $\{\mathbf{x}(\cdot), \mathbf{u}(\cdot), \boldsymbol{\lambda}(\cdot)\}$ and a multiplier $\boldsymbol{\nu}_f$ that satisfy the differential-algebraic equations (DAE),

$$\dot{\mathbf{x}} = \frac{\partial H}{\partial \boldsymbol{\lambda}} \tag{15}$$

$$\dot{\boldsymbol{\lambda}} = -\frac{\partial H}{\partial \mathbf{x}} \tag{16}$$

$$\frac{\partial H}{\partial \mathbf{u}} = \mathbf{0} \tag{17}$$

and the boundary conditions

$$\mathbf{x}(\tau_0) = \mathbf{x}_0 \tag{18}$$

$$\mathbf{e}\big(\mathbf{x}(\tau_f)\big) = \mathbf{0} \tag{19}$$

$$\boldsymbol{\lambda}(\tau_f) = \left(\frac{\partial \mathbf{e}}{\partial \mathbf{x}(\tau_f)}\right)^T \boldsymbol{\nu}_f \tag{20}$$

This is widely recognized as a difficult problem.[6] Thus, although the Minimum Principle is a very useful theory and an extremely important tool in analysis, it is not a numerical method and does not "solve" the problem. Except in rare cases no analytic solutions can be obtained for the BVP. Even linear BVPs do not have analytic solutions. Thus, solving even simple trajectory optimization problems involves numerical methods. Note that the transformation of Problem B to Problem B^λ is not a numerical method. By definition, ***numerical methods involve discretization***. Hence, the type of discretization may be used to classify numerical methods for trajectory optimization. Deferring a discussion on the various types of discretizations, we note that under certain conditions other transformations on Problem B may be possible.

Assume that $\mathbf{u}$ may be eliminated *analytically* from Eq.(2) so that we can write,

$$\mathbf{u}(\tau) = \mathbf{g}(\dot{\mathbf{x}}(\tau), \mathbf{x}(\tau), \tau) \tag{21}$$

Note that this is possible in extremely rare cases. If $\mathbf{f}(\cdot)$ satisfies the conditions for the Inverse Mapping Theorem, then $\mathbf{g}(\cdot)$ exists, but it may not be possible to find it analytically. Whether or not an analytic expression for $\mathbf{g}$ can be obtained, this transformation is the well-known ***hodograph transformation*** frequently used in engineering analysis.[7] Then, substituting Eq.(21) back in Eq.(2), the dynamic constraints transform as

$$\dot{\mathbf{x}}(\tau) = \mathbf{f}(\mathbf{x}(\tau), \mathbf{g}(\dot{\mathbf{x}}(\tau), \mathbf{x}(\tau), \tau), \tau) \tag{22}$$

If this does not result in a trivial equation ($\dot{\mathbf{x}} = \dot{\mathbf{x}}$) then Eq.(22) may be written in the implicit form,

$$\mathbf{f}_{cov}(\mathbf{x}(\tau), \dot{\mathbf{x}}(\tau), \tau) = \mathbf{0} \tag{23}$$

If Eq.(22) reduces to a trivial expression, then the system is said to be differentially flat and is further discussed later in this section. In the same fashion, the running cost, F transforms to F_{cov} and Problem B may be ***transformed*** to a classical calculus of variations problem[4,8] by substituting for the controls in Eqs.(1) and (2). Thus, we get

Problem B_{COV}

Determine the state function, $\mathbf{x}(\cdot)$ that minimizes the fixed-time Lagrange cost functional,

$$J[\mathbf{x}(\cdot)] = \int_{\tau_0}^{\tau_f} F_{cov}(\mathbf{x}(\tau), \dot{\mathbf{x}}(\tau), \tau) d\tau \tag{24}$$

subject to the side constraints,

$$\mathbf{f}_{cov}(\mathbf{x}(\tau), \dot{\mathbf{x}}(\tau), \tau) = \mathbf{0} \tag{25}$$

and the same end-point constraints as before. Note that unlike the transformation of Problem B to B^λ, this transformation (i.e. from B to B_{COV}) is still

in the primal space; consequently, it has a primal-dual formulation given by the classical Euler-Lagrange equations. We denote this problem by Problem B^{λ}_{COV} which is also a BVP.

Now suppose that Problem B is expanded to include simple control constraints, $\mathbf{u} \in \mathcal{U} = \{\mathbf{u} : \mathbf{u}_l \leq \mathbf{u} \leq \mathbf{u}_u\}$. Let us denote this as Problem $B + U$. In other words Problem $B + U$ = Problem B + simple control constraints. Assume again that Problem $B + U$ continues to have the rare feature of allowing an analytic hodograph transformation given by Eq.(21); then, the control constraints, $\mathbf{u} \in \mathcal{U}$, results in

$$\mathbf{u}_l \leq \mathbf{g}(\dot{\mathbf{x}}(\tau), \mathbf{x}(\tau), \tau) \leq \mathbf{u}_u \tag{26}$$

and the dynamic constraints in Problem B_{COV} generalize to,

$$\mathbf{f}_l \leq \mathbf{f}_{cov}(\dot{\mathbf{x}}(\tau), \mathbf{x}(\tau), \tau) \leq \mathbf{f}_u \tag{27}$$

Since Eq.(27) is, strictly speaking, not an equation but a relation, it is representative of the more general relation (see Eq.(9)) known as a ***differential inclusion***. Hence, when Problem B includes a control constraint, the calculus-of-variations problem is modified by Problem B_{DI},

Problem B_{DI}

Determine the state function, $\mathbf{x}(\cdot)$ that minimizes the fixed-time Lagrange cost functional,

$$J[\mathbf{x}(\cdot)] = \int_{\tau_0}^{\tau_f} F_{cov}(\mathbf{x}(\tau), \dot{\mathbf{x}}(\tau), \tau) d\tau \tag{28}$$

subject to the side constraints,

$$\mathbf{f}_l \leq \mathbf{f}_{cov}(\dot{\mathbf{x}}(\tau), \mathbf{x}(\tau), \tau) \leq \mathbf{f}_u \tag{29}$$

and the same end-point constraints as before. Thus, this "differential inclusion problem" is really Problem B_{COV} modified by replacing Eq.(25) with Eq.(27). As with Problem B_{COV}, this is a transformation in the primal space; hence it has a dual counterpart given by the ***adjoint inclusion***.[5] This is a generalized BVP and is denoted by Problem B^{λ}_{DI}. A ***controlled differential inclusion*** – a further generalization of a differential inclusion – is given in Eq.(9). Thus, differential inclusions is no more a method for solving trajectory optimization problems as Problem G is: it is clearly a statement of a problem; it is not a method – numerical or analytical. Unfortunately, it has been erroneously described as a (numerical) method by some practitioners thus creating vast misuse of good terminology. Sometimes, it is also referred to as an ***inverse method***[3,9] since the controls are computed "inversely" from Eq.(21). The process may be generalized by including a coordinate transformation $\mathbf{y} = \mathbf{c}(\mathbf{x})$ in which case $\mathbf{y}$ is referred to as an output and the system so obtained is called an inverse system. This procedure is also referred to as ***inverse dynamics*** or ***dynamic inversion***.[9]

While the concept of differential inclusions has been around since the 1960s,[10] the concept of ***differential flatness*** is relatively new and has gained some exposure in the last few years. According to Fliess et al.,[11] an autonomous dynamical system, (i.e. Eq.(2) with time removed) is said to be differentially flat if there exists an output,

$$\mathbf{y} = \mathbf{c}(\mathbf{x}, \mathbf{u}, \dot{\mathbf{u}}, \ldots, \mathbf{u}^{(\alpha)}), \quad \mathbf{y} \in \mathbb{R}^{N_u} \tag{30}$$

such that the state and controls can be written as

$$\mathbf{x} = \mathbf{a}(\mathbf{y}, \dot{\mathbf{y}}, \ldots, \mathbf{y}^{(\beta)}) \tag{31}$$

$$\mathbf{u} = \mathbf{b}(\mathbf{y}, \dot{\mathbf{y}}, \ldots, \mathbf{y}^{(\beta+1)}) \tag{32}$$

The output, $\mathbf{y}$ is called a ***flat output***. Thus, intuitively, a dynamical system is differentially flat if it is equivalent to a system without dynamics, i.e. a static system. That is, in output space, there are no differential constraints. However, the boundary conditions transform nonlinearly in a possibly complex form according to,

$$\mathbf{x}(\tau_0) = \mathbf{a}(\mathbf{y}(\tau_0), \dot{\mathbf{y}}(\tau_0), \ldots, \mathbf{y}_{(}\tau_0)^{(\beta)}) = \mathbf{x}_0 \tag{33}$$

$$\mathbf{e}\big(\mathbf{x}(\tau_f)\big) = \mathbf{e}\big(\mathbf{a}(\mathbf{y}(\tau_f), \dot{\mathbf{y}}(\tau_f), \ldots, \mathbf{y}_{(}\tau_f)^{(\beta)})\big) = \mathbf{0} \tag{34}$$

Similarly, the running cost, F also transforms in a possibly complex manner. Thus Problem B transforms to a rather standard problem of the calculus-of-variations with some nonlinear boundary conditions. Hence we have the following problem,

Problem B_{DF}

Determine the flat output, $\mathbf{y}(\cdot)$, that minimizes the fixed-time Lagrange cost functional,

$$J[\mathbf{y}(\cdot)] = \int_{\tau_0}^{\tau_f} F_{df}(\mathbf{y}(\tau), \dot{\mathbf{y}}(\tau), \ldots, \mathbf{y}^{(\beta+1)}) d\tau \tag{35}$$

subject to the end point constraints,

$$\mathbf{a}(\mathbf{y}(\tau_0), \dot{\mathbf{y}}(\tau_0), \ldots, \mathbf{y}_{(}\tau_0)^{(\beta)}) = \mathbf{x}_0 \tag{36}$$

$$\mathbf{e}\big(\mathbf{a}(\mathbf{y}(\tau_f), \dot{\mathbf{y}}(\tau_f), \ldots, \mathbf{y}_{(}\tau_f)^{(\beta)})\big) = \mathbf{0} \tag{37}$$

where F_{df} is the transformed cost function. Comparing this problem to Problem B_{COV}, it is clear that if Eq.(22) were to reduce to an identity, then $\mathbf{y} = \mathbf{x}$ is the flat output. Similar to Problem B_{COV}, this also has a dual counterpart, Problem, B^{λ}_{DF}. The apparent advantage of a differentially flat system is that every trajectory in the output space is feasible; therefore, trajectory generation is theoretically simpler in terms of the flat outputs. But, the disadvantages are that it is still difficult to determine whether a given system is differentially flat; consequently, only a limited number of systems can be characterized as differentially flat. Even when systems are flat, solving the problem in output space may generate substantial numerical difficulties since the transformed boundary conditions and the cost function may have undesirable numerical properties.

Further Issues

It is evident that we can distinguish two clear types of transformations: ones that occur in the primal space and another that "lifts" each problem to the primal-dual space. The dualized problems are natural in the sense that to each problem in the primal space, there is a corresponding primal-dual problem facilitated by the Legendre-Fenchel transform. The transformations that occur in the primal space require additional assumptions on the system (e.g. differential flatness). In other cases the transformations may involve a *loss of information*. As noted by Sussmann,[12] "... the passage from the control system to the differential inclusion often involves a loss of information." If one considers Problem B to be limiting (from a point of view of solving complex problems), then the primal transformations of this problem are further limited. None of the primal transformations posed above can be generalized to Problem G. Hence one may regard these transformations in the primal space being applicable to a limited class of problems.

3 Discretizations

As noted earlier, except in very rare cases, none of the problems posed above (either the original problem or the dualized versions) can be solved analytically. By its very nature, a numerical method automatically implies a discrete approximation. The continuous problems of the previous section can be discretely approximated using a variety of methods such as Euler or Runge-Kutta methods. Many engineers tacitly assume a fourth or fifth-order Runge-Kutta method to be the "truth" and often forget that it is still an approximation method. A ***discretization method*** is said to be ***direct*** or ***indirect*** when it refers to the discretization of Problem $B_{(\cdot)}$ or $B^{\lambda}_{(\cdot)}$ respectively. The type of discretization then qualifies the method. Thus, for example, the terms *direct Runge-Kutta* and *indirect Runge-Kutta* are used to denote Runge-Kutta discretization methods for Problem $B_{(\cdot)}$ and $B^{\lambda}_{(\cdot)}$ respectively. Just as the transformation method discussed in the previous section does not "solve" the problem, a discretization method does not solve the problem. Instead it converts infinite dimensional problems to finite dimensional ones.

When Problem $B_{(\cdot)}$ is discretized, the infinite dimensional problem (of finding optimal functions in function space) reduces to the finite dimensional problem of parameter optimization. Thus a discrete version of a continuous nonlinear trajectory optimization Problem $B_{(\cdot)}$ is a structured nonlinear programming problem (NLP)

Problem $B^N_{(\cdot)}$

$$\text{minimize} \qquad J(\mathbf{q}^N_{(\cdot)}) \tag{38}$$

$$\text{subject to} \qquad \mathbf{q}^N_{(\cdot)} \in \mathcal{Q}^{\mathcal{N}}_{(\cdot)} \tag{39}$$

where N denotes the finiteness of the number of parameters (such as a collection of mesh points), $\mathbf{q}^N_{(\cdot)}$ represents the finite collection of parameters that approximate the appropriate functions described in Problem $B_{(\cdot)}$, and the set $\mathcal{Q}^{\mathcal{N}}_{(\cdot)}$ denotes the totality of all the constraints. It is apparent that the structure of $\mathcal{Q}^{\mathcal{N}}_{(\cdot)}$ corresponds in some way to the structure of the constraints specified by Problem $B^N_{(\cdot)}$. Methods that exploit this structure are at the focus of some of the current research topics.[13,14]

A direct method is said to converge if the solution to Problem $B^N_{(\cdot)}$ approaches the solution to Problem $B_{(\cdot)}$ in some norm as $N \to \infty$. This is stated compactly as,

$$\lim_{N\to\infty} B^N_{(\cdot)} \to B_{(\cdot)} \tag{40}$$

This is referred to as the ***convergence of the discretization*** and is not to be confused with the convergence of the NLP algorithm. Convergence of algorithms is discussed in the next section.

In the same spirit, it is apparent that when Problem $B^{\lambda}_{(\cdot)}$ is discretized, the infinite dimensional problem reduces to a finite-dimensional algebraic problem of solving ***generalized nonlinear equations***. A generalized nonlinear equation is stated as $0 \in f(x)$ and is a generalization of the familiar nonlinear equation, $f(x) = 0$. Thus a discrete version of a BVP Problem $B^{\lambda}_{(\cdot)}$ is a root-finding problem (RFP) of solving nonlinear equations if all constraints can be stipulated as equalities. Since Problem $B^{\lambda}_{(\cdot)}$ involves inequalities (for example, Problem B^{λ}_{DI}), the discrete BVP is an RFP involving generalized equations,

Problem $B^{\lambda N}_{(\cdot)}$

$$\text{find} \qquad \mathbf{q}^{\lambda N}_{(\cdot)} \tag{41}$$

$$\text{such that} \qquad \mathbf{q}^{\lambda N}_{(\cdot)} \in \mathcal{Q}^{\lambda\mathcal{N}}_{(\cdot)} \tag{42}$$

where the symbols have similar meanings as in the direct discretization. Comparing Problem $B^{\lambda N}_{(\cdot)}$ and $B^N_{(\cdot)}$, it is apparent that the root-finding problem may be interpreted as finding a feasible solution for a nonlinear programming problem. Thus it is no surprise that modern methods for solving nonlinear equations are based on nonlinear programming. In fact, generalized equations are now routinely solved using NLP codes. Finally, as with the direct discretization, an indirect method is said to converge if,

$$\lim_{N\to\infty} B^{\lambda N}_{(\cdot)} \to B^{\lambda}_{(\cdot)} \tag{43}$$

A natural question that arises in the discretization process is the nature of the connection between Problem $B^N_{(\cdot)}$ and $B^{\lambda N}_{(\cdot)}$. Since the continuous problems are related (via the Legendre-Fenchel transform), an important question to ask is if the discrete problems are related. This question has both numerical and theoretical consequences as shown presently.

Using our now familiar notation, we can dualize Problem $B^N_{(\cdot)}$ to Problem $B^{N\lambda}_{(\cdot)}$. This dualization is achieved by way of the Lagrange multiplier rule and the Karush-Kuhn-Tucker conditions.[15] It is extremely tempting to equate Problem $B^{N\lambda}_{(\cdot)}$ to Problem $B^{\lambda N}_{(\cdot)}$. In general, this is not true! That is discretization and dualization do not commute. It is also very tempting to presume that dualizing the discrete problem preserves the order of the apparent discretization of the dual variable. In general, this is also not true![16] However, there is a body of numerical methods where the discretization and dualization commute with respect to an appropriate transformation. This is articulated to as the Covector Mapping Principle:

The Covector Mapping Principle

Given a general optimal control Problem G, and a discrete approximation to G denoted by Problem G^N, there exists an order-preserving map between the dual variables corresponding to the dualized Problem $G^{N\lambda}$ and the discretized Problem $G^{\lambda N}$.

As noted above, not all methods satisfy this principle which creates a "gap" as illustrated in Figure 1. Methods that satisfy this principle close this gap and are called ***Complete Methods***. The presently known set of complete methods are Hager's family of Runge-Kutta methods[16] and the Legendre pseudospectral method.[17] While Hager's Runge-Kutta methods provide a nontrivial adjoint transformation for the dual variables, the Legendre pseudospectral method provides a simpler transformation in the sense that it is linear and symmetric. The popular Hermite-Simpson method is not a complete method.[18] Many types of Runge-Kutta methods are not complete either. The notion of a complete method blurs the distinction between a direct and indirect method and facilitates convergence theorems.[16,19]

A direct method can be thought of as discretizing first and then dualizing whereas an indirect method is the process of dualizing first and discretizing afterwards. The former is a preferred method for solving complex problems since it involves substantially less labor. The latter is a preferred method from the point of view of accuracy. By virtue of the Covector Mapping Principle, a complete method allows the commutation of discretization and dualization thereby obviating the notion of a direct or indirect method.

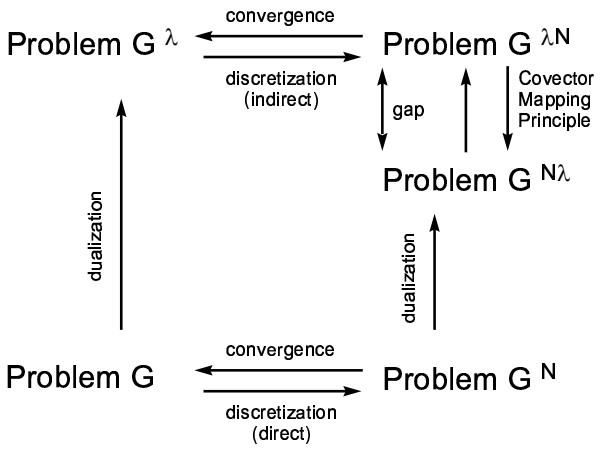

Fig. 1 Schematic of Dualization, Discretization and the Covector Mapping Principle

Open Problems

Covector Mapping Theorems (i.e. special versions of the Covector Mapping Principle) have been proven for a very narrow class of methods. It is further limited to a narrow class of problems (i.e. smooth problems or problems with free end points). Many questions remain unanswered; for example: do the theorems generalize to a larger class of problems, particularly nonsmooth problems where most applications can be categorized? While specific methods can be shown to be complete or otherwise, can necessary conditions be obtained to test whether a method is complete? Given a complete method, how can its corresponding covector mapping theorem be exploited to generate efficient algorithms? Does a complete method automatically imply a convergent method? These are some of the vast number of open questions that are at the heart of theoretical and numerical aspects of solving complex nonlinear optimal control problems. Since answers to these questions facilitate local solutions to the Hamilton-Jacobi equations, it is apparent that optimal feedback control for complex nonlinear systems is possible via this approach.

4 Algorithms

The final step in solving a trajectory optimization problem is an algorithm for solving the discrete problems posed in the previous section. Since Problem $B^N_{(\cdot)}$ is a subset of Problem $B^{\lambda N}_{(\cdot)}$ we will discuss algorithms for solving the structured NLP Problem $B^N_{(\cdot)}$.

An algorithm[15] A is a point-to-set map, $A : \mathbb{A} \rightrightarrows \mathbb{A}$, such that to each point $\mathbf{a}_0 \in \mathbb{A}$, it assigns a sequence, $\{\mathbf{a}_1, \mathbf{a}_2, \ldots\} \in \mathbb{A}$ by way of a point-to-point map, $I : \mathbb{A} \to \mathbb{A}$ called an iteration that generates $\mathbf{a}_{k+1}$ for any given $\mathbf{a}_k, k = 0, 1, \ldots$. A fixed point of I is a point that satisfies $I(\mathbf{a}) = \mathbf{a}$. An algorithm is said to be convergent from a point $\mathbf{a}_0$ if $\mathbf{a} \in A(\mathbf{a}_0)$ where $\mathbf{a}$ is a fixed point of I. Alternatively, an algorithm is said to

be convergent from the point $\mathbf{a}_0$ if,

$$\lim_{k\to\infty} \mathbf{a}_k = \mathbf{a} \quad (= I(\mathbf{a})) \tag{44}$$

From this brief discussion it is clear that we have two notions of convergence with regards to solving trajectory optimization problems: one corresponding to the discretization as described in the previous section and another corresponding to the algorithm defined above. These notions are independent. Thus it is possible to obtain a wrong solution to the trajectory optimization problem by successfully solving Problem $B^N_{(\cdot)}$ (i.e. obtaining an algorithmically convergent solution) by choosing a non-convergent discretization scheme. On the other hand choosing a convergent discretization scheme but a non-convergent algorithm is not as disastrous in the sense that it is a self-correcting mechanism – the lack of a solution!

If the algorithm, A, converges from every point $\mathbf{a}_0 \in \mathbb{A}$, then it is said to be globally convergent.[15] Note that global convergence of the algorithm does not mean it converges to a globally minimum solution. A damped Newton method is globally convergent (under mild conditions) but does not necessarily find a globally minimum solution.

The vector, $\mathbf{d}_k = \mathbf{a}_k - \mathbf{a}_{k-1}$ is the direction of the iteration, I at iterate, k. Various scalar-valued functions, $V : \mathbb{R}^{N_a} \to \mathbb{R}$ are used to define measures of performance of the algorithm such as rate or order of convergence, rate of descent with respect to some descent function and so on. For example the order of convergence of an algorithm at a fixed point $\mathbf{a}$ is defined in terms of some norm function $V_k(\mathbf{a}) = \|\mathbf{a}_k - \mathbf{a}\|$. If $V_k = c|\mathbf{a}_k - \mathbf{a}|^p$ where $|\cdot|$ is the Euclidean norm, c is some constant and $\lim_{k\to\infty} V_k = 0$ then p is called the order of convergence. Thus, if $p = 2$, the rate of convergence is said to be quadratic.

This once again illustrates that the convergence of the algorithm has nothing to do with the convergence of the discretization. The former refers to iterative maps while the latter is with reference to accuracy.

5 Conclusions

An optimal control problem can be transformed to another optimal control problem by way of various transformations. A vast number of methods can simply be categorized as transformations of the problem than as new numerical methods. In fact, such transformations are really a transformation of the coordinates of the dynamical system. When dynamical systems exhibit certain special properties, such transformations may be useful in solving the original problem in the transformed space.

On the other hand, every optimal control problem can be transformed to another problem (not another optimal control problem) by way of the Legendre-Fenchel transform. Hence, an optimal control problem is a primal problem which can be transformed to a primal-dual problem. This transformation is the dualization of the problem and is achieved by way of the Hamiltonian of the problem. Methods that discretize the primal problem are called direct methods while those methods that discretize the transformed (i.e. primal-dual) problem are called indirect. Hence, when a direct method is used to discretize a (primal) problem it indirectly and automatically discretizes its transformed problem, i.e. the primal-dual problem. An important question on the discretization process is whether or not the apparent discretization of the primal-dual problem is of the same order as that of the primal problem. Not all methods preserve the order of the discretization.

The discrete primal problem can also be dualized by the Lagrangian of the problem. One question posed in this paper is whether or not the process of discretizations and dualizations commute. This question is linked to the Covector Mapping Principle. This principle states that discretization and dualization can commute with respect to a map. Such a map exists for Hager's family of Runge-Kutta methods and the Legendre pseudospectral method. It can be shown that such mappings do not exist for certain methods that include a class of Runge-Kutta method and the popular Hermite-Simpson method. Methods for which such maps exist are called complete methods, and by definition, satisfy the covector mapping principle. Proofs of existence of such maps (i.e mapping theorems) are at the heart of new methods of discretization.

The motivation for such new methods for discretization is several-fold. It has been argued that direct methods are not as accurate as indirect methods. Our perspective reveals that this may be attributed to those direct methods that are not complete methods. Thus choosing a complete method provides higher accuracy. Direct methods are preferred over indirect methods, particularly for complex problems, since no labor is required in dualizing the problem; that is, in developing the labor-intensive necessary conditions for optimality. While the labor intensity may be alleviated by symbolic packages, industry-strength problems do not facilitate such luxuries due to various complexities like table-look data that cannot be analytically differentiated. Thus, choosing a complete method provides the advantage of accuracy without the burden of labor. In addition, since a complete method satisfies the covector mapping principle, the associated mapping theorem may be used to determine the all-too-important dual variables. These variables may be use to perform self-checks on the optimality of the result, glean insight on parameter sensitivities, clues on alternative optimas and other quick insights on the nature and structure of the problem. Thus complete methods by virtue of the covector mapping principle essentially blur the distinction between the so-called

direct and indirect methods.

A discretization method is not an algorithm. The performance of many algorithms are based on certain assumptions on the properties of the constraint sets and the functions which are being minimized. Hence, a method must be matched to an appropriate algorithm. Once this matching is accomplished, the algorithm may be tuned to the method by exploiting its properties. A complete method has a distinct property provided by the mapping principle. It is apparent that exploiting this map facilitates a blending of the discretization method to the algorithm. A numerical trajectory optimization method is therefore the totality of the discretization method and the algorithmic method. Too often a method is characterized solely by the discretization method or the algorithmic method. Either characterization is incomplete. Thus, for example, one may use a Sequentially Quadratic Programming (SQP) algorithmic method or say an Interior Point algorithmic method in conjunction with a "direct" Runge-Kutta discretization (collocation) method. Conversely one may use an SQP algorithm to solve the problem by way of shooting or collocation. Naming methods such that they reveal the discretization and the algorithm may generate unusually long names; for example, "Direct Runge-Kutta Collocation Sequential Quadratic Programming Method." Since a discretization method comes prior to the algorithmic method, it seems more appropriate to name a method based on the discretization than the algorithm. This view facilitates the notion of the algorithm being the engine of the discretization method, while the discretization method (along with complementary problems like mesh-refinements) take on the role of the body of the method.

Acknowledgements

We would like to thank Matthew Bottkol of Draper Labs for his insightful mathematical discussions.

References

[1]Betts, J. T., *Practical Methods for Optimal Control Using Nonlinear Programming,* SIAM: Advances in Control and Design Series, Philadelphia, PA, 2001.

[2]Betts, J. T., "Survey of Numerical Methods for Trajectory Optimization," *Journal of Guidance, Control, and Dynamics,* Vol. 21, No. 2, 1998, pp. 193-207.

[3]Bryson, A.E., *Dynamic Optimization,* Addison-Wesley, Menlo Park, CA, 1999.

[4]Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., and Mischenko, E. F., *The Mathematical Theory of Optimal Processes*, Wiley-Interscience, New York, N.Y., 1962.

[5]Clarke, F. H., *Optimization and Nonsmooth Analysis*, SIAM Classics in Applied Mathematics, Philadelphia, PA, 1990.

[6]Ascher, U. M. and Petzold, L. R., *Computer Methods for Ordinary Differential Equations and Differential-Algebriac Equations,* SIAM, Philadelphia, PA, 1998.

[7]Marec, J. P., *Optimal Space Trajectories*, Elsevier Science, 1979.

[8]Bliss, G. A., *Lectures on the Calculus of Variations,* University of Chicago Press, Chicago, IL, 1946.

[9]Nijmeijer, H. and van der Schaft, A. J., *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, N.Y., 1990.

[10]Ioffe, A., "Euler-Lagrange and hamiltonian Formalisms in Dynamic Optimization," *Transactions of the American Mathematical Society*, Vol.349, No.7., 1997, pp.2871-2900.

[11]Fliess, M. Levine, J. Martin, P. and Rouchon, P., "Flatness and Defect of Nonlinear Systems: Introductory Theory and Examples," *International Journal of Control,* Vol.61, No.6, June 1995, pp. 1327-1361.

[12]Sussmann, H. J., "Some Recent Results On The Maximum Principle Of Optimal Control Theory," *Systems and Control in the* 21^{st} *Century*, C. I. Byrnes, B. N. Datta, D. S. Gilliam and C. F. Martin (Eds.), Birkhäuser, Boston, 1997, pp. 351-372.

[13]Barclay, A., Gill, P. E., and Rosen, J. B., "SQP Methods and Their Application to Numerical Optimal Control," Report NA 97-3, Department of Mathematics, University of California, San Diego, La Jolla, CA.

[14]Strizzi, J., Ross, I. M and Fahroo, F., "Towards Real-Time Computation of Optimal Controls for Nonlinear Systems," *Proceedings of the AIAA Guidance, Navigation, and Control Conference,* Monterey, CA, August 2002, AIAA Paper No. 2002-4945.

[15]Bazaraa, M. S., Sherali, H. D., and Shetty, C. M., *Nonlinear Programming: Theory and Algorithms*, John Wiley and Sons, Inc., New York, N.Y., 1993.

[16]Hager, W. W., "Runge-Kutta Methods in Optimal Control and the Transformed Adjoint System," *Numerische Mathematik,* Vol. 87, 2000, pp. 247-282.

[17]Ross, I. M. and Fahroo, F., "A Pseudospectral Transformation of the Covectors of Optimal Control Systems," *Proceedings of the First IFAC Symposium on Systems, Structure and Control*, Prague, The Czech Republic, August 2001.

[18]Enright, P. G. and Conway B. A., "Discrete Approximations to Optimal Trajectories Using Direct Transcription and Nonlinear Programming," *Journal of Guidance, Control, and Dynamics,* Vol. 15, No. 3, 1992, pp. 994-1002.

[19]Ross, I. M. and Fahroo, F.,"Convergence of Pseudospectral Approximations for Optimal Control Problems," *Proceedings of the 2001 IEEE Conference on Decision and Control,*, Orlando, FL, December 2001.