Constructive Controllability Algorithms for Motion Planning and Optimization

W. Todd Cerven
Aeronautical and Astronautical Engineering
University of Illinois at Urbana-Champaign
104 S Wright St, Urbana, IL 61801, USA
Tel: (217) 244-8516
Email: w-cerven@uiuc.edu

Francesco Bullo
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
1308 W Main St, Urbana, IL 61801, USA
Tel: (217) 333-0656, Fax: (217) 244-1653
Email: bullo@uiuc.edu

Abstract

This paper presents novel algorithms for planning feasible and minimum energy paths. The algorithms rely on series expansions to characterize planning problems for polynomial control systems. The resulting inversion problem is solved through an iterative contraction or a power series inversion. While the design methodology is general, our focus is on a class of polynomial control systems for which we provide explicit convergence guarantees. We demonstrate performance and numerical characteristics using one dimensional and six dimensional systems.

I. Introduction

Problem Description

In recent years, the study of autonomous vehicles and versatile robots has become a highly active research area as technological advances have enabled these devices to cross over from science fiction to reality. This development has the capability to impact a variety of areas such as factory automation, search and rescue operations, oceanographic and aerospace missions, and medical robotics, with advances in computing and manufacturing allowing increasing levels of autonomy and dexterity in diverse environments. Such progress requires control systems able to independently plan motions both rapidly and reliably. How to generate the control profiles required to pass from one state to another

An early short version of this work appeared in [1].

is the realm of motion planning and optimal control. Analytic and numerical techniques have been developed to address this challenge.

Lie brackets based planners are another set of algorithms for open loop control design; see [2], [3]. The typical planner relies on oscillations in order to move, in a way similar to how one parks a car or how an animal changes its shape to locomote. These methods have been applied to chained form systems [4], driftless systems [5], [6], locomotion systems [7], [8], and addressed by the authors' earlier work [9], [10]. The classic limitation of Lie bracket methods is their local nature, as only small amplitude motions can be planned satisfactorily.

Differential flatness is an intrinsic property of nonlinear control systems introduced in [11], [12]. For systems enjoying this property motion planning is a straightforward task; e.g., see [13]. However, there are a few limitations in the differential flatness approach to motion planning. First of all, there is no algorithmic procedure to obtain the flat outputs of a system, nor is there any criterion to establish whether flat outputs exist. Second, differential flatness is a non-generic property. In other words, generic variations or perturbations to a flat model of a system render the system "non-flat." This is particularly relevant to the context of vehicle models, where the famous flat PVTOL model is not known to be flat as soon as aerodynamic forces are modeled more accurately (added mass due to movement through fluid [14] or drag terms [13]). Finally, little is known with regards to "approximately flat" systems; see [15]. There is no established notion of approximate flatness, nor is it clear that motion planning would be a straightforward problem for systems enjoying such a property.

In *numerical optimal control*, [16], trajectories are obtained through a numerical optimization. Because the problem is infinite dimensional, various forms of transcription (i.e., discretization or parametrization) are used to cast the variational problem into a nonlinear program. Collocation [17] uses base functions to parameterize both the states and the controls, while differential inclusion [18] avoids using the discretized controls by explicitly solving for the controls in terms of the states and their derivatives. Although useful and powerful, the high dimension, complexity, and lack of convergence guarantees limit the speed and reliability of these algorithms.

Although non-optimal, other techniques based on *heuristics randomization* have been developed that promise fast execution in complex environments. These algorithms focus more on obstacle avoidance than on the nonlinear dynamics of the system. Popular among these are solutions based upon roadmaps [19] and incremental searches [20]. In roadmap methods, path planning is accomplished in two steps: a collection of sample configurations is selected, and then trajectories connecting all sample points are computed. The second step, however, is a local controllability problem in itself.

In this paper, we avoid the limitations currently inherent to flatness, the limitations of Lie bracket-

based planners, and the computational issues of numerical optimization to investigate a method based on power series to quickly generate the required control profiles.

Power series are not new to nonlinear control. They have been used widely, notably in the nonlinear regulator problem. Al'brekht [21] used power series to solve the Hamilton-Jacobi-Bellman (HJB) partial differential equations to obtain an optimal stabilizing control. For the same problem, Halme and coworkers [22], [23], [24], [25] developed a polynomial power series and a local inverse to generalized power series. Krener and coworkers [26], [27] use Al'brekht's method to solve for the nonlinear regulator corresponding to the Francis-Byrnes-Isidori equations. Using level set methods and power series about extremal trajectories they also propose a method to extend Al'brekht's solutions to the HJB equations well beyond the neighborhood of the origin.

Of course, power series methods are invariably local, but that does not make them any less relevant or useful. As mentioned beforehand, randomized methods require local motion planners. Autonomous and semi-autonomous vehicles also need such a method, as large scale maneuvers are relatively easily user defined, but small ones are impractical, if not impossible, for a user to direct. Other applications include station keeping for under-actuated aerospace and underwater vehicles [28], [29] and movement based on internal actuation, swimming, and other biologically motivated designs [30], [28], [31].

Statement of contribution

This paper builds on the aforementioned areas of research to develop local complete constructive trajectory generation and optimization algorithms for a class of low order polynomial systems which is representative of a large array of dynamical systems. These algorithms are complete in that they guarantee a solution and are "constructive" in the sense that they rely directly on the controllability properties of the system.

This paper presents two algorithms to generate a feasible path using base functions and a minimum energy path parameterized by the initial values of the costates of the system. For a linearly controllable system, we can show that there exists a neighborhood about the origin in which the algorithms are guaranteed to find a solution. To find these parameterized controls, we develop iterative as well as series inversion methods, both of which have convergence guarantees. We provide proofs to this along with computation of explicit neighborhoods that, even if conservative, provide a lower bound on region of validity, or the region over which these algorithms are guaranteed to converge. Additionally, we investigate the behavior of the algorithms for a one dimensional system and for a planar vertical takeoff and landing vehicle (PVTOL) with damping. This includes an illustration of the level of conservativeness of the lower bounds on the neighborhoods of convergence for the example systems.

This paper is organized as follows. In Section II, we introduce the polynomial systems of interest and define the norms and series expansions upon which these algorithms are based. We discuss the accessibility and the nilpotency of the polynomial system as well as the formulation required to apply the series expansion about a trajectory. Next, in Section III, we present trajectory generation and optimization problems and we cast both of them into the form of a function inversion problem. In Section IV, we proceed to show how a unique solution to the inversion problem exists locally, and define two numerical approaches to compute it. A lower bound to the radius of convergence is provided for both methods. Lastly, in Section V, we apply these algorithms to a simple one dimensional example and the PVTOL with damping. Appendices I, II, and III contain various proofs.

II. A class of polynomial control systems

Throughout the paper we shall concern ourselves with n-dimensional second order polynomial systems of the form

$$\begin{aligned} \dot{x} &= Ax + f^{[2]}(x,x) + Bu \\ x(0) &= x_0, \end{aligned} \tag{1}$$

where $f^{[2]} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a symmetric tensor,[1] A is an $n \times n$ matrix, and B is an $n \times m$ matrix. While the approach advocated in this work can be extended to address more general systems, we focus on this class of polynomial systems for simplicity of exposition. This class is representative of a large array of dynamical systems, as any smooth system linear in controls not fitting this form naturally can be approximated as such by a Taylor expansion. Classical dynamical systems such as the Lorentz, Lotka-Volterra, and Euler equations are characterized by second order polynomial vector fields. In addition, Kang and Krener [32] showed that any nonlinear system of the form $\dot{\xi} = f(\xi) + g(\xi)\mu$ can be represented as such (plus higher order terms) by a change of coordinates and state feedback. Note that this class of polynomial systems is not contained in the class of systems in chained form, driftless systems, and feedback linearizable systems.

This set of nonlinearities appears in many common mechanical systems as, for example, trigonometric functions can usually be rewritten in polynomial form. Let us consider, for instance, control systems arising from rigid body dynamics. We let the state x comprise the absolute position $q \in \mathbb{R}^3$, the body-fixed translational and angular velocities $v, \omega \in \mathbb{R}^3$, and the orientation matrix $(R + I_3) \in SO(3)$. For various vehicles including aircraft, spacecraft, and watercraft, the equations of motion can be

[1] Any vector field with components homogeneous polynomials of degree 2 can be written in terms of a symmetric tensor $f^{[2]}$.

represented as:

$$\dot{x} = \begin{bmatrix} \dot{q} \\ \dot{R} \\ \dot{v} \\ \dot{\omega} \end{bmatrix} = \begin{bmatrix} Rv \\ R\widehat{\omega} \\ -\widehat{\omega}v + G_v(R,q) + D_v(v) + B_v u \\ -\mathbb{I}^{-1}\widehat{\omega}\mathbb{I}\omega + G_\omega(R,q) + D_\omega(v,\omega) + B_\omega u \end{bmatrix}, \tag{2}$$

where $\mathbb{I}$ is the inertia matrix and B_v and B_ω are constant matrices. The functions G_v and G_ω are force and torque due to gravity, D_v and D_ω are the damping terms. These equations fit the form of (1) either naturally or by truncating the Taylor expansions for the functions G_v, G_ω, D_v and D_ω.

A. Operator and function norms

In defining mapping and norms we follow the notation in [33, Chapter 6].

Let $\mathbb{N}$ be the set of strictly positive integers. Over the linear space $\mathbb{R}^n$ we will use the norms $\|x\|_2 = \sqrt{x'x}$, and $\|x\|_\infty = \max_{i\in\{1,\dots,n\}} |x_i|$. Consider the normed linear space $\mathcal{L}_\infty^n$ of piecewise continuous, uniformly bounded functions over the interval I

$$\begin{aligned} x : I \subset \mathbb{R}_+ &\to \mathbb{R}^n \\ t &\mapsto x(t), \end{aligned}$$

with norm

$$\|x\|_{\mathcal{L}_\infty} = \sup_{t\in I} \|x(t)\|_\infty = \sup_{t\in I} \max_{i\in\{1,\dots,n\}} |x_i(t)| < +\infty.$$

Assume the matrix A is Hurwitz or that the interval I is finite, and let H_A be the mapping

$$\begin{aligned} H_A : \mathcal{L}_\infty^n &\to \mathcal{L}_\infty^n \\ x(t) &\mapsto \int_0^t e^{A(t-\tau)} x(\tau) d\tau. \end{aligned}$$

The $\mathcal{L}_\infty^n$ induced norm for H_A is

$$\|H_A\|_{\mathcal{L}_\infty} = \|e^{At}\|_{\mathcal{L}_1} = \max_{i=1,\dots,n} \sum_{j=1}^n \int_I |(e^{At})_{ij}| dt.$$

Next, we consider 2-tensor $f^{[2]} : \mathcal{L}_\infty^n \times \mathcal{L}_\infty^n \to \mathcal{L}_\infty^n$ defined via

$$(x(t), y(t)) \mapsto f^{[2]}(x(t), y(t)),$$

and define its induced norm $\|f^{[2]}\|_{\mathcal{L}_\infty}$ via

$$\|f^{[2]}\|_{\mathcal{L}_\infty} = \|f^{[2]}\|_\infty = \max_{\substack{\|y_1\|_\infty=1 \\ \|y_2\|_\infty=1}} \|f^{[2]}(y_1, y_2)\|_\infty.$$

B. Evolution as series expansion

We present a series expansion for the solution of the initial value problem in equation (1). The result is an extension of the results in [34] and is proven in Appendix I. Before proceeding, it is useful to introduce a few preliminary concepts. The Catalan numbers are an infinite sequence of integers discovered by Euler as the solution to the question *"How many ways can a convex polygon be divided into triangles via non-intersecting diagonals?"*. The result is a sequence of numbers corresponding to polygons with increasing numbers of vertices, starting with a triangle. We define the Catalan numbers $\{c_k \in \mathbb{R}, k \in \mathbb{N}\}$ as in [35, Section 2.3]; our definition differs from Euler's more standard sequence by a scaling factor. Define $\mathcal{C} : [0,1] \to [0,1]$ as $\mathcal{C}(\eta) = 1 - \sqrt{1-\eta}$, and let $\text{Remainder}_K(\mathcal{C})(\eta)$ be its Taylor remainder of order K. If we develop $\mathcal{C}$ in power series

$$\mathcal{C}(\eta) = \sum_{k=1}^{+\infty} c_k \eta^k,$$

then the following equivalent conditions hold

$$c_k = \frac{1}{k\, 2^{2k-1}} \binom{2k-2}{k-1}, \qquad \text{and} \qquad c_1 = \frac{1}{2}, \;\; c_k = \frac{1}{2} \sum_{i=1}^{k-1} c_i c_{k-i}. \tag{3}$$

We are now able to characterize the flow of the differential equation (1).

Lemma II.1: The solution of the system in equation (1) is written as a series $x(t) = \sum_{k=1}^{+\infty} x_k(t)$ where

$$\begin{aligned} x_1(t) &= e^{At} x_0 + \int_0^t e^{A(t-\tau)} B u(\tau) d\tau \\ x_k(t) &= \sum_{a=1}^{k-1} \int_0^t e^{A(t-\tau)} f^{[2]}(x_a(\tau), x_{k-a}(\tau)) d\tau, \quad \forall k > 1. \end{aligned} \tag{4}$$

Let $d_1 = 2\left(\|e^{At}x_0\|_{\mathcal{L}_\infty} + \|e^{At}\|_{\mathcal{L}_1}\|Bu\|_{\mathcal{L}_\infty}\right)$ and $d_2 = 2\|e^{At}\|_{\mathcal{L}_1}\|f^{[2]}\|_{\mathcal{L}_\infty}$. Provided $d_1 d_2 \leq 1$, a solution exists over the interval I and the series converges absolutely and uniformly in $t \in I$, and the following upper bounds hold:

$$\|x_k\|_{\mathcal{L}_\infty} \leq c_k\, d_1^k d_2^{k-1},$$

$$\|x - \sum_{k=1}^{K} x_k\|_{\mathcal{L}_\infty} \leq \frac{1}{d_2} \text{Remainder}_K(\mathcal{C})(d_1 d_2).$$

C. Accessibility and nilpotency

Consider the polynomial control system in equation (1) described by the tensors A, B, and $f^{[2]}$. Let the subspace $\mathcal{B} \subset \mathbb{R}^n$ be the image of the matrix B. Given two linear subspaces V and W of $\mathbb{R}^n$, let

$$f^{[2]}(V, W) = \{f^{[2]}(v, w) \in \mathbb{R}^n |\; v \in V, w \in W\} \subset \mathbb{R}^n.$$

Let $\text{LinReach}_A(\mathcal{B})$ be the subspace generated by

$$B, AB, \ldots, A^n B,$$

and define the *accessibility subspaces* $\{\mathcal{R}_k \subset \mathbb{R}^n, k \in \mathbb{N}\}$ as follows

$$\begin{aligned}
\mathcal{R}_1 &= \text{LinReach}_A(\mathcal{B}) = \text{span}\{B, AB, \ldots, A^n B\} \\
\mathcal{R}_2 &= \text{LinReach}_A(f^{[2]}(\mathcal{R}_1, \mathcal{R}_1)) \\
&\vdots \\
\mathcal{R}_k &= \text{LinReach}_A(\cup_{a=1}^{k-1}\{f^{[2]}(\mathcal{R}_a, \mathcal{R}_{k-a})\}).
\end{aligned}$$

The subspaces $\{\mathcal{R}_k \subset \mathbb{R}^n, k \in \mathbb{N}\}$ play a key role in studying controllability and nilpotency of system (1). In particular, we state the following facts:
(i) the kth order component $x_k(t)$ is in $\mathcal{R}_k$ for all $t \in I$ and for all inputs $u : I \to \mathbb{R}^m$,
(ii) the system is *linearly controllable* if and only if $\mathcal{R}_1$ is full rank,
(iii) the accessibility subspace $\mathcal{R}_k$ is generated by all the Lie brackets evaluated at the origin — between an arbitrary number of the vector field $Ax + f^{[2]}(x, x)$ and k vector fields of the form B_i,
(iv) the system is *locally accessible* if $\sum_{k=1}^{+\infty} \mathcal{R}_k = \mathbb{R}^n$, and
(v) the system is *nilpotent* if there exists an integer k such that $\mathcal{R}_i = 0$ for all $i \geq k$.

D. Series expansion about a trajectory

As described in Lemma II.1 the series expansion in equation (4) converges under the assumption of small initial condition $x(0)$. There is a second setting in which a similar expansion can be easily written. Assume that a reference trajectory satisfying $\dot{x} = Ax + f^{[2]}(x, x) + Bu_{\text{ref}}(t)$, $x(0) = x_0$ is known analytically as $x(t) = \Phi_t^f(x_0, u_{\text{ref}}(t))$. Define a relative variation variable e

$$e = x - \Phi_t^f(x_0, u_{\text{ref}}(t)),$$

and compute the differential equation regulating its evolution

$$\begin{aligned}
\dot{e} &= \dot{x} - \dot{\Phi}_t^f(x_0, u_{\text{ref}}(t)) \\
&= \left(Ax + f^{[2]}(x, x) + Bu\right) - \left(A\Phi_t^f(x_0, u_{\text{ref}}(t)) + f^{[2]}(\Phi_t^f(x_0, u_{\text{ref}}(t)), \Phi_t^f(x_0, u_{\text{ref}}(t))) + Bu_{\text{ref}}(t)\right) \\
&= Ae + f^{[2]}(x + \Phi_t^f(x_0, u_{\text{ref}}(t)), x - \Phi_t^f(x_0, u_{\text{ref}}(t))) + B(u - u_{\text{ref}}(t)) \\
&= Ae + f^{[2]}(2\Phi_t^f(x_0, u_{\text{ref}}(t)) + e, e) + B(u - u_{\text{ref}}(t)) \\
&= \left(A + 2f^{[2]}(\Phi_t^f(x_0, u_{\text{ref}}(t)))\right) e + f^{[2]}(e, e) + B(u - u_{\text{ref}}(t)),
\end{aligned}$$

where we define the matrix $f^{[2]}(x)$ according to $\left(f^{[2]}(x)\right) y = f^{[2]}(x, y)$ for all $x, y \in \mathbb{R}^n$.

In the new variable e, the system is again in second order polynomial form and the initial condition is $e(0) = 0$. The time-varying nature of the matrix $A + 2f^{[2]}(\Phi_t^f(x_0, u_{\text{ref}}(t)))$ is the only difference between this system and the system in equation (1). One can show that the series expansion in equation (4) can be written for a time-varying system by replacing the kernel e^{At} in the convolution integrals with the more general state transition matrix for time-varying linear systems. Furthermore, under certain technical conditions, the equations are once again time-invariant as in (1). One such example is given by the setting of rigid body vehicles (2) evolving along helical trajectories, see [36].

III. Formulation of motion planning and minimum energy planning problems

This section describes two interesting planning problems. We transform these problems into inverse function problems exploiting the series expansion described above.

Consider the control system in equation (1), let the initial condition be the origin $x(0) = 0$, and let $x_d \in \mathbb{R}^n$ be the desired target location. We shall require that $d_1 d_2 \leq 1$, i.e., we restrict our investigation to the convergence radius of the series in equation (4).

A. Base functions for the control inputs

It is often useful to introduce a collection of bounded piecewise continuous base functions $\{\psi^i(t) : [0, T] \mapsto R^m, i \in \{1, \ldots, n_p\}\}$ to parametrize the input functions u. This is the case for example when magnitude and rate constraints or binary actuators are present. We write

$$u(t) = \sum_{i=1}^{n_p} \psi^i(t) p_i = \psi(t) p.$$

A wide variety of base functions are possible including splines, Hermite polynomials, sinusoidal functions, piecewise constant functions, and wavelets. Define the tensors $\{\Phi_k : \mathbb{R}^{k \times n_p} \to \mathbb{R}^n, k \in \mathbb{N}\}$ as:

$$\begin{aligned}
\Phi_1^i(t) &= \int_0^t e^{A(t-\tau)} B \psi^i(\tau) d\tau \\
\Phi_2^{i_1 i_2}(t) &= \int_0^t e^{A(t-\tau)} f^{[2]}\left(\Phi_1^{i_1}(\tau), \Phi_1^{i_2}(\tau)\right) d\tau, \\
\Phi_3^{i_1 i_2 i_3}(t) &= \int_0^t e^{A(t-\tau)} \Big(f^{[2]}(\Phi_1^{i_1}(\tau), \Phi_2^{i_2 i_3}(\tau)) + f^{[2]}(\Phi_2^{i_1 i_2}(\tau), \Phi_1^{i_3}(\tau)) \Big) d\tau \\
&\vdots \\
\Phi_k^{i_1 \ldots i_k}(t) &= \sum_{a=1}^{k-1} \int_0^t e^{A(t-\tau)} f^{[2]}(\Phi_a^{i_1 \ldots i_a}(\tau), \Phi_{k-a}^{i_1 \ldots i_{k-1}}(\tau)) d\tau.
\end{aligned} \tag{5}$$

Assuming $x(0) = 0$, the kth term of the series in equation (4) can now be rewritten as

$$x_k(t) = \Phi_k(t)(\underbrace{p, \ldots, p}_{k \text{ times}}).$$

In what follows, we will only need $\Phi_k(t)$ evaluated at final time T, therefore we introduce the abbreviation $\Phi_k = \Phi_k(T)$.

B. Motion planning with base functions

Consider the following design problem: find a control input $u : [0, T] \mapsto R^m$ such that

$$\dot{x} = Ax + f^{[2]}(x, x) + Bu$$
$$x(0) = 0, \;\; x(T) = x_{\text{target}}.$$

Using the series expansion characterization in Section II-B, the problem becomes finding a control input $u : [0, T] \mapsto R^m$ that solves

$$x_{\text{target}} = \sum_{k=1}^{+\infty} x_k(T).$$

This equation is a constraint on the input functions u since all the terms x_k depend on it. This constraint can be discretized into a finite dimensional equation via a collection of bounded piecewise continuous base functions $\{\psi^i(t) : [0, T] \mapsto R^m, i \in \{1, \ldots, n_p\}\}$. Using the notation in Section III-A, the design problem is to find a vector $p \in \mathbb{R}^{n_p}$ such that

$$x_{\text{target}} = \Phi_1 p + \sum_{k=2}^{+\infty} \Phi_k(p, \ldots, p). \tag{6}$$

C. Minimum energy planning (without base functions)

Consider the following design problem: find a control input $u : [0, T] \mapsto R^m$ that solves

$$\begin{aligned} \min \;\; & \int_0^T \|u(t)\|_2^2 dt \\ \text{subject to } \; & \dot{x} = Ax + f^{[2]}(x, x) + Bu \\ & x(0) = 0, \;\; x(T) = x_{\text{target}}. \end{aligned} \tag{7}$$

Thus, the Hamiltonian associated with the optimal control problem in equation (7) is:

$$H(x, \lambda, u) = \frac{1}{2}\|u\|_2^2 + \lambda' \left(Ax + f^{[2]}(x, x) + Bu\right).$$

As known from optimal control theory, we let u extremize H, that is, we let $u(t) = -B'\lambda(t)$, where B' denotes the transpose of B and we write necessary conditions

$$\begin{aligned}\dot{x} &= Ax + f^{[2]}(x,x) - BB'\lambda \\ \dot{\lambda} &= -A'\lambda - 2f^{[2]}(x)'\lambda \\ x(0) &= 0, \;\; x(T) = x_{\text{target}}.\end{aligned} \tag{8}$$

The design problem is to find the initial value $\lambda(0) = \lambda_0$ compatible with problem (8) that uniquely determines the optimal control law.

The two point boundary value problem has the same polynomial structure of the initial value problem in equation (1). We let $\overline{x} = (x, \lambda) \in \mathbb{R}^{2n}$, and

$$\overline{A} = \begin{bmatrix} A & -BB' \\ 0 & -A' \end{bmatrix}, \quad \overline{f}^{[2]}(\overline{x}, \overline{x}) = \begin{bmatrix} f^{[2]}(x,x) \\ -2f^{[2]}(x)'\lambda \end{bmatrix},$$

so that the first order term in the solution to (8) is

$$\overline{x}_1(t) = \begin{bmatrix} x_1(t) \\ \lambda_1(t) \end{bmatrix} = e^{\overline{A}t} \begin{bmatrix} 0 \\ \lambda_0 \end{bmatrix} = \overline{\Phi}_1(t) \begin{bmatrix} 0 \\ \lambda_0 \end{bmatrix},$$

where $\overline{\Phi}_1$ now maps $\mathbb{R}^{2n} \to \mathbb{R}^{2n}$. The higher order terms $\{\overline{\Phi}_k : \mathbb{R}^{k \times 2n} \to \mathbb{R}^{2n}, k > 1\}$ can be recursively defined following equation (5) in Section III-A. Using the series expansion characterization in Section II-B, we have

$$\begin{bmatrix} x(T) \\ \lambda(T) \end{bmatrix} = \overline{\Phi}_1 \begin{bmatrix} 0 \\ \lambda_0 \end{bmatrix} + \sum_{k=2}^{+\infty} \overline{\Phi}_k \left(\begin{bmatrix} 0 \\ \lambda_0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ \lambda_0 \end{bmatrix} \right),$$

where we drop the T argument as usual. This expression can be rewritten as

$$x_{\text{target}} = \overline{\Phi}_{1,x\lambda}\lambda_0 + \sum_{k=2}^{+\infty} \overline{\Phi}_{k,x\lambda} \left(\lambda_0, \ldots, \lambda_0\right), \tag{9}$$

where we project the image and restrict the domain of the tensor $\{\overline{\Phi}_k, k \in \mathbb{N}\}$ as

$$\overline{\Phi}_k \left(\begin{bmatrix} 0 \\ \lambda_0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ \lambda_0 \end{bmatrix} \right) = \left(\begin{bmatrix} \overline{\Phi}_{k,x\lambda} \left(\lambda_0, \ldots, \lambda_0\right) \\ \overline{\Phi}_{k,\lambda\lambda} \left(\lambda_0, \ldots, \lambda_0\right) \end{bmatrix} \right).$$

In summary, the design problem is to find a vector $\lambda_0 \in \mathbb{R}^n$ solution to equation (9). Once an appropriate value of λ_0 is found, the optimal control law can be computed as a series expansion.

IV. Solving the planning problems via inversion

In this section, we treat both motion planning and minimum energy planning as a function inversion problem for an appropriate function f characterized via a power series. We study conditions that

guarantee that the function f and its Jacobian are invertible. Finally, we describe two approaches to inverting f and to bound the neighborhood over which the function is invertible.

Equations (6) and (9) are equivalent to the solution of an equation of the form

$$x_{\text{target}} = f(y) = f_1 y + \sum_{k=2}^{+\infty} f_k(y, \dots, y), \tag{10}$$

where $y = p \in \mathbb{R}^{n_p}$ for motion planning and $y = \lambda_0 \in \mathbb{R}^n$ for minimum energy planning. Additionally, $x_{\text{target}} \in \mathbb{R}^n$, and the tensors f_k live in linear spaces of appropriate dimensions. Next, we transcribe the bounds known from Lemma II.1 into the new setting. Let the sequence $\{c_k, k \in \mathbb{N}\}$ and the function $\mathcal{C}$ be defined as in Section II-B.

Lemma IV.1: Define

$$D_1 = \begin{cases} 2\|e^{At}\|_{\mathcal{L}_1}\|B\psi(t)\|_{\mathcal{L}_\infty} \\ 2\|e^{\overline{A}t}\|_{\mathcal{L}_\infty} \end{cases} \quad \text{and} \quad D_2 = \begin{cases} 2\|e^{At}\|_{\mathcal{L}_1}\|f^{[2]}\|_{\mathcal{L}_\infty} \\ 2\|e^{\overline{A}t}\|_{\mathcal{L}_1}\|\overline{f}^{[2]}\|_{\mathcal{L}_\infty} \end{cases}.$$

Provided $D_1 D_2 \|y\|_\infty \leq 1$, the series converges absolutely, and the following upper bounds hold:

$$\|f_k(y, \dots, y)\|_\infty \leq c_k \, D_1^k D_2^{k-1} \|y\|_\infty^k,$$
$$\|f(y) - \sum_{k=1}^{K} f_k(y, \dots, y)\|_\infty \leq \frac{1}{D_2} \operatorname{Remainder}_K(\mathcal{C}) \left(D_1 D_2 \|y\|_\infty\right).$$

Proof: We relate the coefficients $\{d_1, d_2\}$ for each settings to $\{D_1, D_2\}$ via

$$d_1 = 2\left(\|e^{At}x_0\|_{\mathcal{L}_\infty} + \|e^{At}\|_{\mathcal{L}_1}\|Bu\|_{\mathcal{L}_\infty}\right) = \begin{cases} 2\|e^{At}\|_{\mathcal{L}_1}\|B\psi(t)c\|_{\mathcal{L}_\infty} \\ 2\left\| e^{\overline{A}t} \begin{bmatrix} 0 \\ \lambda_0 \end{bmatrix} \right\|_{\mathcal{L}_\infty} \end{cases} \leq D_1 \|y\|_\infty$$

$$d_2 = D_2.$$

From Lemma II.1 we transcribe the first of the two bounds

$$\|f_k(y, \dots, y)\|_\infty \leq \begin{cases} \|\Phi_k(c, \dots, c)\|_\infty \\ \|\overline{\Phi}_k(\lambda_0, \dots, \lambda_0)\|_\infty \end{cases} = \begin{cases} \|x_k\|_\infty \\ \|\overline{x}_k\|_\infty \end{cases} \leq c_k \, d_1^k d_2^{k-1} \leq c_k \, D_1^k D_2^{k-1} \|y\|_\infty^k.$$

The second bound can be proven using the definition of remainder:

$$\begin{aligned} \|f(y) - \sum_{k=1}^{K} f_k(y, \dots, y)\|_\infty &\leq \frac{1}{d_2} \operatorname{Remainder}_K(\mathcal{C})(d_1 d_2) \\ &\leq \frac{1}{d_2} \sum_{k=K+1}^{+\infty} c_k (d_1 d_2)^k \leq \frac{1}{D_2} \sum_{k=K+1}^{+\infty} c_k \left(D_1 D_2 \|y\|_\infty\right)^k \\ &\leq \frac{1}{D_2} \operatorname{Remainder}_K(\mathcal{C}) \left(D_1 D_2 \|y\|_\infty\right). \end{aligned}$$

■

Theorem IV.2 (A generalized inverse function theorem) Assume the power series in equation (10) converges absolutely over $V_a = \{y \in \mathbb{R}^{n_p} |\ D_1 D_2 \|y\|_\infty \leq 1\}$, and let $f : \mathbb{R}^{n_p} \to \mathbb{R}^n$, $n_p \geq n$ be the function defined by the series. Furthermore, assume that the tensor f_1 is full rank. Then there exists a neighborhood $V_b \subseteq V_a$ such that, for all $x_{\text{target}} \in f(V_b)$, there exists a smooth right inverse $f^{-1} : f(V_b) \to V_b$. Furthermore, if $n_p = n$, f^{-1} is unique.

Proof: It can be seen that $f_1 = \frac{\partial f}{\partial y}(0)$ is the Jacobian of f evaluated at $y = 0$. Since f_1 is full rank, we can compute its pseudo-inverse f_1^p. Let $\chi \in \mathbb{R}^n$ and let $y = f_1^p \chi$. Then equation (10) becomes

$$x_{\text{target}} = \chi + \sum_{k=2}^{+\infty} f_k(f_1^p \chi, \ldots, f_1^p \chi) = h(\chi).$$

The Jacobian of the function $h : \mathbb{R}^n \to \mathbb{R}^n$ evaluated at the origin is

$$\frac{\partial h}{\partial \chi}(0) = I_n.$$

Therefore, the function h has a unique inverse in a neighborhood of the origin because of Theorem 2.5.2 in [37]. This implies that f^{-1} exists in a neighborhood of the origin. Furthermore, when $n_p = n$, the inverse function f^{-1} is unique since the pseudo-inverse f_1^p equals f_1^{-1}. ■

A. Existence of solution for linearly controllable systems

Motivated by the previous theorem, we investigate necessary and sufficient conditions in order for the tensor f_1 to be full rank. It turns out that in both settings the property of linear controllability plays a central role.

Lemma IV.3: There exist smooth base functions $\{\psi^i(t) : i \in \{1, \ldots, n_p\}\}$ such that the tensor Φ_1 is invertible if and only if the system in equation (1) is linearly controllable.

Proof: If the tensor Φ_1 is full rank, then the linear systems obtained by setting $f^{[2]}$ to zero is controllable, and therefore the system in equation (1) is linearly controllable. Vice-versa, let $n_p = n$ and define the functions

$$\begin{aligned} \psi^i(t) &= B' e^{A'(T-t)} W_T^{-1} \mathsf{e}_i \\ W_T &= \int_0^T e^{A(T-s)} B B' e^{A'(T-s)} ds = W_T' > 0, \end{aligned} \tag{11}$$

where $\{\mathsf{e}_1, \ldots, \mathsf{e}_n\}$ is the standard base for $\mathbb{R}^n$ and W_T is the controllability Grammian. As this system is linearly controllable by assumption, W_T is full rank. Given these input base functions, it is easy to see that $\Phi_1 = I_n$. ■

The base functions in equation (11) are selected according to the classic *minimum energy design* for point to point planning of linear control systems; see [38, page 557].

Lemma IV.4: The tensor $\overline{\Phi}_{1,x\lambda}$ is invertible if and only if the system in equation (1) is linearly controllable.

Proof: The tensor $\overline{\Phi}_{1,x\lambda}$ can be found by solving the differential equation

$$\dot{\overline{\Phi}}_1 = \begin{bmatrix} \dot{\overline{\Phi}}_{1,xx} & \dot{\overline{\Phi}}_{1,x\lambda} \\ \dot{\overline{\Phi}}_{1,\lambda x} & \dot{\overline{\Phi}}_{1,\lambda\lambda} \end{bmatrix} = \begin{bmatrix} A & -BB' \\ 0 & -A' \end{bmatrix} \begin{bmatrix} \overline{\Phi}_{1,xx} & \overline{\Phi}_{1,x\lambda} \\ \overline{\Phi}_{1,\lambda x} & \overline{\Phi}_{1,\lambda\lambda} \end{bmatrix}, \quad \overline{\Phi}_i(0) = I_n,$$

which simplifies to

$$\dot{\overline{\Phi}}_{1,x\lambda} = A\overline{\Phi}_{1,x\lambda} - BB'e^{-A't}, \quad \overline{\Phi}_{1,x\lambda}(0) = 0_n.$$

The solution to the last equation is the convolution integral

$$\overline{\Phi}_{1,x\lambda} = -e^{AT}\left(\int_0^T e^{-As}BB'e^{-A's}ds\right).$$

Since $\overline{\Phi}_{1,x\lambda}$ is the negative of the product of an invertible matrix e^{AT} and the controllability Grammian of the system (A, B), $\overline{\Phi}_{1,x\lambda}$ is full rank and invertible if and only if the system (A, B) is controllable. ■

B. Existence of solution for linearly uncontrollable systems

Now let us consider systems that are not linearly controllable.

Theorem IV.5: Given the n-dimensional time invariant dynamical equation (1), if its linear controllability matrix has rank $n_c < n$, then there exists a transformation $\widetilde{x} = Px$, where P is a constant nonsingular matrix, which transforms (1) into

$$\dot{\widetilde{x}} = \begin{bmatrix} \widetilde{A}_c & \widetilde{A}_{12} \\ \widetilde{A}_{nc} & 0 \end{bmatrix} \widetilde{x} + \widetilde{f}^{[2]}(\widetilde{x}, \widetilde{x}) + \begin{bmatrix} \widetilde{B}_c \\ 0 \end{bmatrix} u \tag{12}$$

with the controllable n_c-dimensional subsystem

$$\dot{\widetilde{x}}_c = \widetilde{A}_c\widetilde{x}_c + \widetilde{B}_c u. \tag{13}$$

This transformation is called the system's *canonical decomposition*; see [38]. The matrix P can be defined such that P^{-1} is composed of first n_c independent columns of the controllability matrix $[B \quad AB \ldots A^{n-1}B]$ augmented by arbitrary vectors that make the matrix nonsingular. The linear controllability Grammian $\widetilde{W}_t$ for the canonical decomposition (12) reduces to

$$\begin{bmatrix} \widetilde{W}_{t_c} & 0 \\ 0 & 0_{n-n_c} \end{bmatrix}, \tag{14}$$

where $\widetilde{W}_{t_c} = \int_0^t e^{\widetilde{A}_c(t-s)}\widetilde{B}_c\widetilde{B}_c' e^{\widetilde{A}_c'(t-s)}ds$ is the controllability Grammian for the system (13). Using the canonical decomposition, the inversion problem (10) can be recast as

$$\begin{bmatrix} \widetilde{x}_{c_{\text{target}}} \\ \widetilde{x}_{nc_{\text{target}}} \end{bmatrix} = \widetilde{f}(\widetilde{y}) = \begin{bmatrix} \widetilde{f}_{c_1} \\ \widetilde{f}_{nc_1} \end{bmatrix} \widetilde{y} + \sum_{k=2}^{+\infty} \begin{bmatrix} \widetilde{f}_{c_k} \\ \widetilde{f}_{nc_k} \end{bmatrix} (\widetilde{y}, \ldots, \widetilde{y}), \tag{15}$$

where $\widetilde{x}_{c_{\text{target}}} \in \mathbb{R}^{n_c}$ and $\widetilde{x}_{nc_{\text{target}}} \in \mathbb{R}^{n-n_c}$. Let us then ignore motion on the linearly uncontrollable space and reduce our planning problem to that on the linearly controllable space $\mathbb{R}^{n_c}$

$$\widetilde{x}_{c_{\text{target}}} = \widetilde{f}_c(\widetilde{y}_c) = \widetilde{f}_{c_1}\widetilde{y}_c + \sum_{k=2}^{+\infty} \widetilde{f}_{c_k}(\widetilde{y}_c, \ldots, \widetilde{y}_c), \tag{16}$$

where $\widetilde{y} = [\widetilde{y}_c' \quad \widetilde{y}_{nc}']' \in \mathbb{R}^{n_p}$ and $\widetilde{y}_c \in \mathbb{R}^{n_c}$.

This problem is now in a form where the inverse function theorem can be applied, so we again investigate necessary and sufficient conditions in order for the tensor $\widetilde{f}_{c_1}$ to be full rank.

Lemma IV.6: There exist smooth base functions $\{\psi^i(t) : i \in \{1, \ldots, n_p\}\}$ such that the tensor $\widetilde{f}_{c_1} = \widetilde{\Phi}_{c_1}$ is invertible if and only if the system in equation (1) is linearly controllable on the space $\mathbb{R}^{n_c}$.

Proof: This proof follows that of Lemma IV.3, where $n_p = n$ and the base functions are defined as

$$\psi^i(t) = \widetilde{B}' e^{\widetilde{A}'(T-t)} \begin{bmatrix} \widetilde{W}_{T_c}^{-1} & 0 \\ 0 & 0_{n-n_c} \end{bmatrix} \mathsf{e}_i,$$

where $\{\mathsf{e}_1, \ldots, e_n\}$ is the standard base for $\mathbb{R}^n$. It is then straightforward to find that

$$\widetilde{f}_1 = \begin{bmatrix} I_{n_c} & 0 \\ 0 & 0_{n-n_c} \end{bmatrix}.$$

and $\widetilde{f}_{c_1} = I_{n_c}$ is invertible. ■

Lemma IV.7: The tensor $\widetilde{f}_{c_1} = \widetilde{\overline{\Phi}}_{c_{1,x\lambda}}$ is full rank and is invertible if and only if the system in equation (1) is linearly controllable on the space $\mathbb{R}^{n_c}$.

Proof: This proof follows that of Lemma IV.4. The linear term $\widetilde{f}_1$ is then defined as

$$\widetilde{f}_1 = \widetilde{\overline{\Phi}}_{1,x\lambda} = -e^{\widetilde{A}T}\left(\int_0^T e^{-\widetilde{A}s}\widetilde{B}\widetilde{B}' e^{-\widetilde{A}'s}ds\right) = \begin{bmatrix} \widetilde{W}_{T_c} & 0 \\ 0 & 0_{n-n_c} \end{bmatrix}$$

It can then be seen that $\widetilde{y} = \widetilde{\lambda} \in \mathbb{R}^n$ and $\widetilde{f}_{c_1} = \widetilde{W}_{T_c}$ has inverse $(\widetilde{W}_{T_c}^{-1})$ if and only if the system (12) is linearly controllable. ■

Remark IV.8: The treatment of linearly uncontrollable systems is particularly important when considering nonminimum or redundant coordinate representations. Here, a *nonminimum coordinate representation* is a coordinate parametrization of a configuration space for which the number of coordinates

exceeds the dimension of the space. Such representations are often important to avoid singularities and write certain dynamical system in quadratic form (1). For example, unit quaternions or rotation matrices are very common to model planar and 3D orientations. Furthermore, they are naturally associated to quadratic vector fields; see the discussion on the model in equation (2) and the PVTOL with damping example below.

C. Computational approaches

This section presents two methodologies to solve the inverse function problem under the linear controllability assumption. First, we note that equation (10) can be solved numerically by a root-finding method such as the classic Newton's method. This type of routine is well-documented in books such as [39] and its implementation is relatively straightforward. Along these lines, we present here a provably convergent iterative method based upon the contraction mapping. We provide an explicit lower bound on the region of convergence of the algorithm. Second, we provide an explicit inverse function written in power series expansion. The closed form expressions here are taken from [23], [24], [25]. Again, we provide an explicit lower bound on the region of convergence of the algorithm.

Iterative contraction algorithm

Define the pseudo-inverse f_1^p and let $y = f_1^p \chi$, where χ is the new free variable living in $\mathbb{R}^n$. We rewrite equation (10) into the equivalent expression:

$$x_{\text{target}} = \chi + \sum_{k=2}^{+\infty} f_k(f_1^p \chi, \ldots, f_1^p \chi). \tag{17}$$

Define the map $\mathcal{M} : \mathbb{R}^n \to \mathbb{R}^n$

$$\mathcal{M}(\chi) = x_{\text{target}} - \sum_{k=2}^{+\infty} f_k(f_1^p \chi, \ldots, f_1^p \chi),$$

and set up the iteration

$$\begin{aligned} \chi_1 &= x_{\text{target}} \\ \chi_{n+1} &= x_{\text{target}} - \sum_{k=2}^{+\infty} f_k(f_1^p \chi_n, \ldots, f_1^p \chi_n) = \mathcal{M}(\chi_n). \end{aligned}$$

We shall prove convergence of this iteration starting from any initial condition inside the set

$$S = \{\chi : \|\chi - x_{\text{target}}\|_\infty \leq \|x_{\text{target}}\|_\infty\}.$$

Theorem IV.9: If

$$\|x_{\text{target}}\|_\infty < \Lambda_1 = \frac{1}{2\|f_1^p\|_\infty D_1 D_2} \min\left\{ \frac{1}{D_1\|f_1^p\|_\infty}, 1 - \frac{(D_1\|f_1^p\|_\infty)^2}{(1+D_1\|f_1^p\|_\infty)^2} \right\},$$

there exists a unique χ^* belonging to the set S and satisfying $\chi^* = \mathcal{M}(\chi^*)$. Furthermore, the unique solution can be computed by iterating the map $\mathcal{M}$ starting from any initial condition in S.

The proof to this theorem can be found in Appendix II. By this theorem, the set V_b in Theorem IV.2 contains a ball of radius Λ_1 about the origin.

Power series inversion algorithm

Next, we present an explicit inverse to the function. We borrow the result from [24], [25]. Consider the power series in equation (10)

$$x_{\text{target}} = f(y) = f_1 y + \sum_{k=2}^{+\infty} f_k(y, \ldots, y).$$

Let $m = n$, and assume that f_1 is invertible. Define the function $g : \mathbb{R}^n \mapsto \mathbb{R}^n$ via the power series

$$g(x) = g_1 x + \sum_{k=2}^{+\infty} g_k(x, \ldots, x), \tag{18}$$

where we let

$$g_1 = f_1^{-1}, \qquad g_k(x, \ldots, x) = -g_1 \sum_{m=2}^{k} \sum_{\substack{i_1+\cdots+i_l=k \\ i_1,\cdots,i_l<k}} f_l\Big(g_{i_1}(x, \cdots, x), \cdots, g_{i_l}(x, \cdots, x)\Big).$$

Theorem IV.10: The function g is the inverse of f, and it converges provided

$$\|x_{\text{target}}\|_\infty \leq \Lambda_2 = \frac{1}{4(D_1\|f_1^p\|_\infty + 1)\|f_1^p\|_\infty D_1 D_2} < \Lambda_1.$$

The proof to this theorem can be found in Appendix III. By this theorem, V_b in Theorem IV.2 contains a ball of radius Λ_2 about the origin.

V. Simulation

Two models were used to illustrate the algorithms. First, a one dimensional nonlinear system $\dot{x} = -x^2 + u$ was used to study the convergence properties of these algorithms. Second, a planar vertical takeoff and landing aircraft model was chosen to test the minimum-energy planning algorithm performance on a more complicated system.

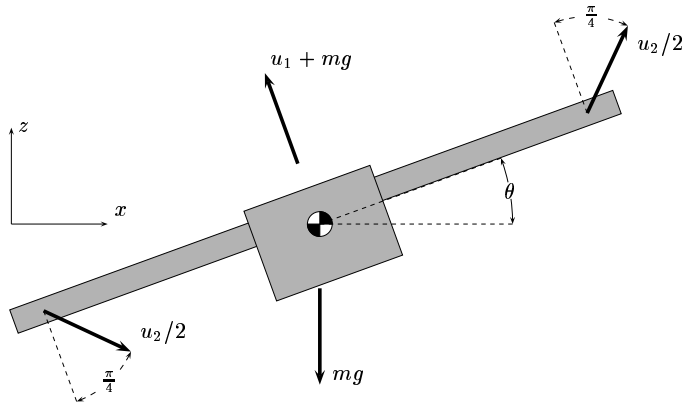

Fig. 1. Diagram of the PVTOL model.

A. PVTOL with Damping Example

We consider the model of a simple planar vertical takeoff and landing aircraft model based upon that of [40], [41], but with added viscous damping forces; see Figure 1. In other words, we consider the classic PVTOL model subject to a linear drag force. We parameterize its configuration and velocity space via the state variables $(s, c, x, z, \omega, v_x, v_z)$. We let x and z be the inertial coordinates of the aircraft and s and c represent its roll angle θ such that $s = \sin\theta$ and $c = \cos\theta - 1$. The angular velocity is ω and the linear velocities in the body-fixed x (respectively z) axis are v_x (respectively v_z). Explicitly separating the linear from the homogeneous polynomial component, the equations are written as

$$\begin{bmatrix} \dot{s} \\ \dot{c} \\ \dot{x} \\ \dot{z} \\ \dot{\omega} \\ \dot{v}_x \\ \dot{v}_z \end{bmatrix} = \begin{bmatrix} \omega \\ 0 \\ v_x \\ v_z \\ \frac{-k_1}{J}\omega \\ \frac{-k_2}{M}v_x - gs \\ \frac{-k_3}{M}v_z - gc \end{bmatrix} + \begin{bmatrix} c\omega \\ -s\omega \\ cv_x - sv_z \\ sv_x + cv_z \\ 0 \\ v_z\omega \\ -v_x\omega \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & \frac{h}{J}k_u \\ 0 & \frac{1}{M}k_u \\ \frac{1}{M}k_u & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

As stated in Section II, the quadratic term can be represented via a symmetric tensor $f^{[2]}$. In components, let us write the ith component of $f^{[2]}(x, x)$ as $(f^{[2]})_i^{jk} x_j x_k$. All components of $f^{[2]}$ vanish except for $(f^{[2]})_i^{jk} = (f^{[2]})_i^{kj} = 1/2$ at indices $(1, 2, 5), (3, 2, 6), (4, 1, 6), (4, 2, 7), (6, 5, 7)$ and $(f^{[2]})_i^{jk} = (f^{[2]})_i^{kj} = -1/2$ at indices $(2, 1, 5), (3, 1, 7), (7, 5, 6)$. The control u_1 corresponds to the body vertical force minus gravity, while u_2 corresponds to coupled forces on the wingtips with a net horizontal component. The other forces depend upon the constants k_i, which parameterize some damping force, and g, the gravity constant. The constant h is the distance from the center of mass to the

wingtip, while M and J are mass and moment of inertia, respectively. The constant k_u is a control gain.

Remark V.1: Although motion planning for the classic PVTOL example has been done effectively using flatness [41], the PVTOL with damping model appears not to be flat. A system with state x and control u is differentially flat if there exists an output function $\vartheta(x, u, \dot{u}, \ddot{u}, \dots)$ such that the states and controls can be expressed solely as a function of the output and its derivatives. The PVTOL equations with damping can be rewritten as follows:

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \frac{M}{k_u} \begin{bmatrix} -\sin\theta & \cos\theta \\ \cos\theta & \sin\theta \end{bmatrix} \left(\begin{bmatrix} \ddot{x} \\ \ddot{z} \end{bmatrix} + \begin{bmatrix} \frac{k_2}{M}\cos^2\theta + \frac{k_3}{M}\sin^2\theta & (\frac{k_2}{M} - \frac{k_3}{M})\sin\theta\cos\theta \\ (\frac{k_2}{M} - \frac{k_3}{M})\sin\theta\cos\theta & \frac{k_2}{M}\sin^2\theta + \frac{k_3}{M}\cos^2\theta \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{z} \end{bmatrix} + \begin{bmatrix} 0 \\ g \end{bmatrix} \right) \tag{19}$$
$$u_2 = \frac{J}{k_u h}\left(\ddot{\theta} + \frac{k_1}{J}\dot{\theta}\right).$$

Equating the first and third equations, we obtain

$$\frac{J}{h}\left(\ddot{\theta} + \frac{k_1}{J}\dot{\theta}\right) - M\left(\ddot{x}\cos\theta + \ddot{z}\sin\theta + \frac{k_2}{M}\left(\dot{x}\cos\theta + \dot{z}\sin\theta\right) + g\sin\theta\right) = 0. \tag{20}$$

For the classical PVTOL (when the damping coefficients are zero), the flat output is

$$\vartheta = \begin{bmatrix} x \\ z \end{bmatrix} + \frac{J}{hM}\begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}.$$

This is also known as the *Huygens center of oscillation.* Inserting the flat output into (20), the angle θ is found [42], [41] to be related to the flat output via $\ddot{\vartheta}_1\cos\theta + (\ddot{\vartheta}_2 + g)\sin\theta = 0$. Once θ is derived, the states and controls can then be calculated by using the output relation and equations of motion, respectively. However, when the damping coefficients are nonzero,

$$(\ddot{\vartheta}_1 + \frac{k_2}{M}\dot{\vartheta}_1)\cos\theta + (\ddot{\vartheta}_2 + \frac{k_2}{M}\dot{\vartheta}_2 + g)\sin\theta = \frac{J}{hM}\left(\frac{k_1}{J} + \frac{k_2}{M}\right)\dot{\theta}.$$

Thus, θ can no longer be written in terms of ϑ and its derivatives, so that the classical PVTOL flat output is no longer a flat output of the system with damping. It is unclear whether flat output still exist.

B. Implementation

The two algorithms were divided into two implementation steps: preprocessing and control derivation. Preprocessing includes the system definition and the calculation of the corresponding tensors in the series expansion. The resulting expansion can be saved to memory for use by the control derivation. The control derivation includes solution of the inverse problem using the contraction method

and the calculation of the controls with respect to that solution. The contraction method was chosen because it both has a larger lower bound on its radius of convergence as well as a more straightforward implementation. The simulation was carried out by numerical solution of the ordinary differential equations. Each of these tasks was implemented in Maple 5.4, primarily due to the nontrivial nature of the calculation of the required tensors. As this involves computation of a series of tensors of increasing dimension, each defined by lower order tensors, it necessitates a data type with expandable structure. This is not straightforward in programming languages such as C, nor in numerical software such as Matlab. Another disadvantage of Matlab is that its tensor manipulation routines are not as comprehensive as its matrix routines, thus requiring nested loops to carry out tensor calculations. While Maple is less computationally efficient than either of the aforementioned methods (documentation [43] suggests floating point computations in Maple can be 50 to 500 times slower than in equivalent Fortran programs), its tensor package accommodates tensor products as well as calculation of the tensors using the recursive functions, avoiding data structure issues.[2] Another computational challenge was posed by the PVTOL system itself. Its controllability Grammian is ill-conditioned (using the parameters described below, its condition number [44, page 56] is in the order of $1e+5$), thus requiring careful treatment and high accuracy. Fortunately, these issues take place in the preprocessing stage and can be tackled offline. These tensor calculations dominate the preprocessing and require, at most, $O(n_{tot}^{k+3})$ multiplications and integration of $O(n_{tot}^{2k+4})$ terms, where n_{tot} and k are the total dimension[3] of the system and the order of the series expansion, respectively (assuming $n \geq k > 1$). The integration then proves to be the primary factor in run time. The control derivation is far less computationally intensive, as it involves primarily floating point computations. Yet, because the number of recursions needed to find a solution for a given accuracy is variable, the number of online calculations is less predictable. This, too, was implemented in Maple, although any programming language would work as well. For the PVTOL example, using a second order series approximation, on an 800 Mhz Windows ME PC using 128 megabytes of RAM, the algorithm took 98.5 seconds in preprocessing and 2.8 seconds (7 iterations) in solving for the control. Third order series calculations took $13,173$ seconds in preprocessing and 20.9 seconds (23 iterations) in solving for the control online, corresponding to the computational estimate above. All of the necessary series data stored for the control derivation stage amounted to 27 and 168 kilobytes for the second and third order expansions, respectively.

[2] An implementation in Mathematica was found to encounter similar features as in Maple.

[3] For the base function algorithm, $n_{tot} = n$, while $n_{tot} = 2n$ for the minimum energy algorithm

C. Results

C.1 Convergence Study

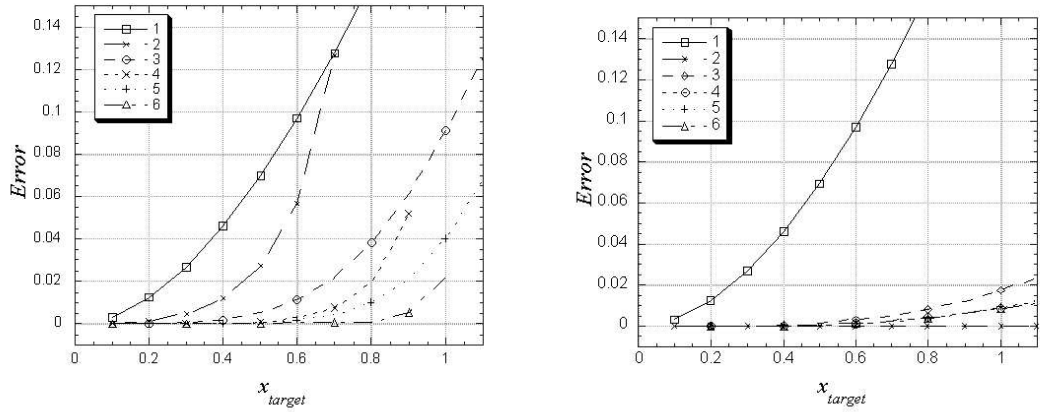

Fig. 2. Final position error for motion planning algorithm (left), minimum energy planning algorithm (right).

The one dimensional system was used to show the solution convergence properties of both algorithms. For the motion planning algorithm, the inverse problem simplifies to the following root finding problem and control definition:

$$x_{\text{target}} = p - \frac{1}{3}\,p^2 + \frac{2}{15}\,p^3 - \frac{16}{315}\,p^4 + \frac{58}{2835}\,p^5 - \frac{1262}{155925}\,p^6 + \ldots, \quad u = p.$$

The lower bound of the neighborhood of convergence of the algorithm, as defined in Theorem IV.9, is $\|x_{\text{target}}\|_\infty < .0625$. As the control is a constant, the solution to the differential equation, for a positive coefficient p, can be written as $x = \sqrt{p}\,\tanh(\sqrt{p}\,t)$. Truncating the series at orders one through six, the corresponding coefficients and controls were found. Figure 2 shows the comparative error among the levels of truncation for a range of x_{target}. This shows a general decrease in error as the order of the truncation increases. The x_{target} at which the even truncated series cease to have a solution corresponds to the maximums of the truncated polynomials. It can therefore be seen that the actual convergence radius of the algorithm is orders of magnitude greater than the minimum described in Theorem IV.9.

For the minimum energy planning algorithm, the inverse problem simplifies to the following root finding problem and control definition:

$$\begin{aligned} x_{\text{target}} &= -\lambda_0 + 0\,{\lambda_0}^2 - \frac{1}{10}\,{\lambda_0}^3 + 0\,{\lambda_0}^4 - \frac{1}{180}\,{\lambda_0}^5 + 0\,{\lambda_0}^6 + \ldots \\ u &= -\lambda_0 + t^2{\lambda_0}^2 - \frac{1}{2}\,t^4{\lambda_0}^3 + \frac{1}{10}\,t^6{\lambda_0}^4 - \frac{1}{20}\,t^8{\lambda_0}^5 + \frac{7}{450}\,t^{10}{\lambda_0}^6 + \ldots\,. \end{aligned}$$

The lower bound of the neighborhood of convergence of the algorithm, as defined in Theorem IV.9, is $\|x_{\text{target}}\|_\infty < .0023$. The control input is computed via a series expansion on the initial value of the costate λ_0. Figure 2 shows the comparative error among the levels of truncation for a range of x_{target}. This shows a general decrease in error as the order of the truncation increases, although this is not true uniformly. This is not unexpected, as the error reflects the accuracy of the solution of x only, ignoring λ. For example, the second order expansion solves the differential equation and constraints on x exactly, but does not solve as accurately for λ. Thus, a feasible trajectory is generated that is suboptimal. Despite this apparent non-uniformity, the algorithm behaves very well at x_{target}, orders of magnitude beyond the conservative minimum provided by Theorem IV.9.

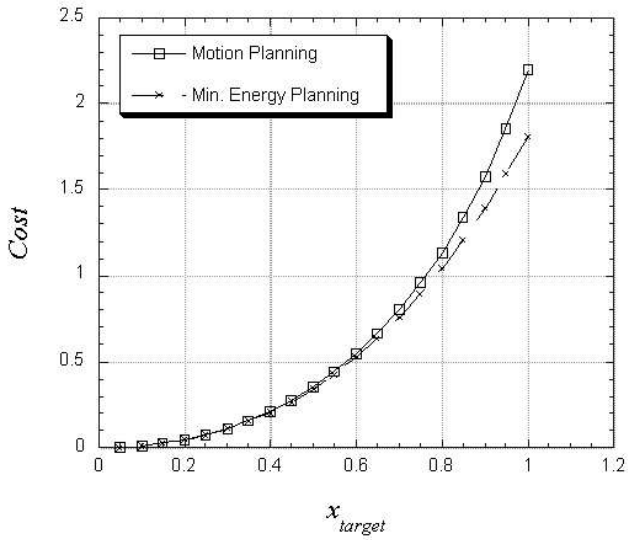

Fig. 3. Cost comparison between motion planning with base functions and minimum energy planning algorithms for series truncated at order 6.

Figure 3 provides a cost comparison between the two techniques for series truncated at order 6. Understandably, as the target distance increases, the control is active longer and the cost differential is more apparent, with a difference of 18% of the optimal cost at $x_{\text{target}} = 1$.

Figure 4 compares the state histories of the linear case as well as the motion and minimum energy planning algorithms of order six. As in the previous figures, the optimal algorithm consistently reaches x_{target} with greater accuracy. Both methods significantly outperform their linear counterpart.

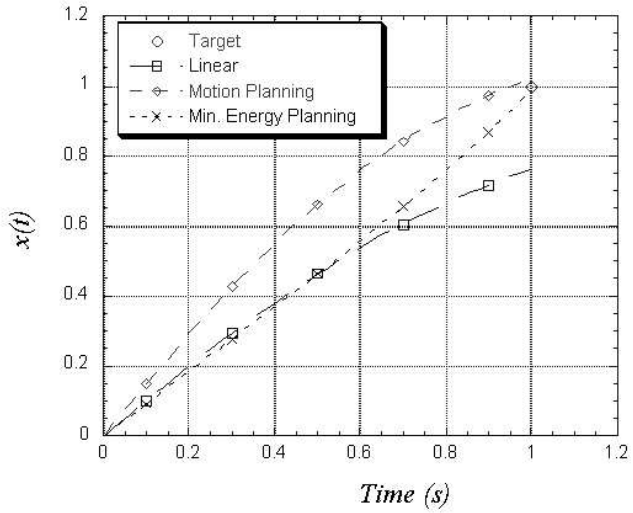

Fig. 4. Trajectory comparison of the motion planning with base functions and minimum energy planning algorithms for series truncated at order 6 with their linear counterpart.

C.2 PVTOL

The minimum energy planning algorithm, having showed good performance for the one dimensional case, was applied to that of the PVTOL. For this case, the aforementioned model was chosen with the constants $k_1 = J = k_2 = k_3 = m = k_u = 50$, $h = 1$, and $g = 10$. As defined in Theorem IV.9, the lower bound of the neighborhood of convergence of the algorithm is $\|x_{\text{target}}\|_\infty < 1.6e-13$. As with the one dimensional case, this was over-conservative, as solutions could be found over $1e+10$ times greater than the bound. One such example is shown in Figure 5, where a positive x displacement of 0.005 was requested with an x component of velocity of -0.0005. Using the series truncated at second order shows a clear improvement over the linear solution, as the error in the final second order state is negligible in comparison to the 3 percent error in the final position of the linear case.

VI. Conclusions

We have presented a variety of constructive controllability and minimum energy control algorithms. The results are local in nature but constructive: existence, uniqueness and optimality are guaranteed for a class of polynomial systems. Lower bounds on the region of validity of these algorithms are presented, and evaluated with respect to algorithm performance in specific examples. As for future research, we plan to investigate how to extend the region of validity of the algorithms by combining

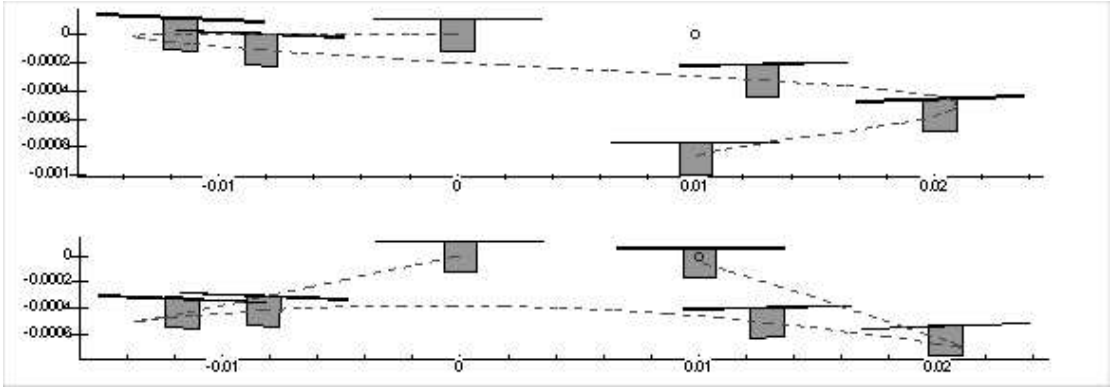

Fig. 5. Resulting motions from minimum energy planning algorithm for PVTOL series truncated at orders 1 (top) and 2 (bottom), starting at rest at the origin with desired final condition of a small negative velocity at the circled location.

them with randomized planning methods [19], [20] and level set methods [27].

References

[1] F. Bullo and W. T. Cerven, "On trajectory optimization for polynomial systems via series expansions," in *IEEE Conf. on Decision and Control*, (Sydney, Australia), pp. 772–777, Dec. 2000.

[2] R. W. Brockett, "Control theory and singular Riemannian geometry," in *New Directions in Applied Mathematics* (P. Hilton and G. Young, eds.), (New York, NY), pp. 11–27, Springer Verlag, 1982.

[3] R. W. Brockett, "Asymptotic stability and feedback stabilization," in *Geometric Control Theory* (R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds.), (Boston, MA), pp. 181–191, Birkhäuser, 1983.

[4] R. M. Murray and S. S. Sastry, "Nonholonomic motion planning: Steering using sinusoids," *IEEE Transactions on Automatic Control*, vol. 5, no. 38, pp. 700–726, 1993.

[5] H. J. Sussmann, "New differential geometric methods in nonholonomic path finding," in *Systems, Models, and Feedback: Theory and Applications* (A. Isidori and T. J. Tarn, eds.), pp. 365–384, Boston, MA: Birkhäuser, 1992.

[6] N. E. Leonard and P. S. Krishnaprasad, "Motion control of drift-free, left-invariant systems on Lie groups," *IEEE Transactions on Automatic Control*, vol. 40, no. 9, pp. 1539–1554, 1995.

[7] J. P. Ostrowski, "Steering for a class of dynamic nonholonomic systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 8, pp. 1492–1497, 2000.

[8] H. Zhang and J. P. Ostrowski, "Control algorithms using affine connections on principal fiber bundles," in *IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control*, (Princeton, NJ), pp. 129–34, Mar. 2000.

[9] F. Bullo, N. E. Leonard, and A. D. Lewis, "Controllability and motion algorithms for underactuated Lagrangian systems on Lie groups," *IEEE Transactions on Automatic Control*, vol. 45, no. 8, pp. 1437–1454, 2000.

[10] W. T. Cerven and V. L. Coverstone-Carroll, "Optimal reorientation of multibody spacecraft through joint motion using averaging theory," *AIAA Journal of Guidance, Control, and Dynamics*, vol. 24, no. 4, pp. 788–795, 2001.

[11] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, "Flatness and defect of non-linear systems: Introductory theory and examples," *International Journal of Control*, vol. 61, no. 6, pp. 1327–1361, 1995.

[12] P. Martin and P. Rouchon, "Any (controllable) driftless system with 3 inputs and 5 states is flat," *Systems & Control Letters*, vol. 25, no. 3, pp. 167–173, 1995.

[13] M. J. van Nieuwstadt and R. M. Murray, "Rapid hover to forward flight transitions for a thrust vectored aircraft," *AIAA Journal of Guidance, Control, and Dynamics*, vol. 21, no. 1, pp. 93–100, 1998.

[14] M. Rathinam and R. M. Murray, "Configuration flatness of Lagrangian systems underactuated by one control," in *IEEE Conf. on Decision and Control*, (Kobe, Japan), pp. 378–389, Dec. 1996.

[15] A. Banaszuk and J. Hauser, "Approximate feedback linearization: A homotopy operator approach," *SIAM Journal on Control and Optimization*, vol. 34, no. 5, pp. 1533–1554, 1996.

[16] A. E. Bryson and Y.-C. Ho, *Applied Optimal Control: Optimization, Estimation, and Control*. Bristol, PA: Taylor & Francis, 1981.

[17] C. R. Hargraves and S. W. Paris, "Direct trajectory optimization using nonlinear programming and collocation," *AIAA Journal of Guidance, Control, and Dynamics*, vol. 10, no. 4, pp. 338–342, 1987.

[18] H. Seywald, "Trajectory optimization based on differential inclusion," *AIAA Journal of Guidance, Control, and Dynamics*, vol. 17, no. 3, pp. 480–487, 1994.

[19] L. E. Kavraki, P. Švestka, J. C. Latombe, and M. H. Overmars, "Probabilistic roadmaps for path planning in high-dimensional space," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 4, pp. 566–580, 1996.

[20] S. M. LaValle and J. J. Kuffner, "Randomized kinodynamic planning," *International Journal of Robotics Research*, vol. 20, no. 5, pp. 378–400, 2001.

[21] E. G. Al'brekht, "On the optimal stabilization of nonlinear systems," *PMM - Journal of Applied Mathematics and Mechanics*, vol. 25, pp. 1254–1266, 1961.

[22] A. Halme, "Polynomial operators for nonlinear systems analysis," *Acta Polytechnica Scandinavica*, vol. Ma24, pp. 7–63, 1972.

[23] A. Halme and J. Orava, "Generalized polynomial operators for nonlinear systems analysis," *IEEE Transactions on Automatic Control*, vol. 17, no. 2, pp. 226–8, 1972.

[24] J. Orava and A. Halme, "Inversion of generalized power series representations," *Journal of Mathematical Analysis and Applications*, vol. 45, pp. 136–141, 1974.

[25] R. P. Hamalainen and A. Halme, "A solution of nonlinear TPBVP's occurring in optimal control," *Automatica*, vol. 12, no. 5, pp. 403–15, 1976.

[26] A. J. Krener, "The construction of optimal linear and nonlinear regulators," in *Systems, Models and Feedback: Theory and Applications* (A. Isidori and T. J. Tarn, eds.), pp. 301–322, Boston, MA: Birkhäuser, 1992.

[27] C. L. Navasca and A. J. Krener, "Solution of Hamilton-Jacobi-Bellman equations," in *IEEE Conf. on Decision and Control*, (Sydney, Australia), pp. 570–574, Dec. 2000.

[28] H. Singh, D. Yoerger, and A. Bradley, "Issues in AUV design and deployment for oceanographic research," in *IEEE Int. Conf. on Robotics and Automation*, (Albuquerque, NM), pp. 1857–1862, Apr. 1997.

[29] C. Rui, I. V. Kolmanovsky, and N. H. McClamroch, "Nonlinear attitude and shape control of spacecraft with articulated appendages and reaction wheels," *IEEE Transactions on Automatic Control*, vol. 45, no. 8, pp. 1455–69, 2000.

[30] K. A. Mclsaac and J. P. Ostrowski, "A geometric approach to anguilliform locomotion: modelling of an underwater eel robot," in *IEEE Int. Conf. on Robotics and Automation*, (Detroit, MI), pp. 2843–8, May 1999.

[31] K. A. Morgansen, V. Duidam, R. Mason, J. W. Burdick, and R. M. Murray, "Nonlinear control methods for planar carangiform robot fish locomotion," in *IEEE Int. Conf. on Robotics and Automation*, (Seoul, Korea), pp. 427–434, Apr. 2001.

[32] W. Kang and A. J. Krener, "Extended quadratic controller form and dynamic state feedback linearization of nonlinear systems," *SIAM Journal on Control and Optimization*, vol. 30, no. 6, pp. 1319–1337, 1992.

[33] H. K. Khalil, *Nonlinear Systems*. Englewood Cliffs, NJ: Prentice Hall, second ed., 1995.

[34] F. Bullo, "Series expansions for analytic systems linear in controls," *Automatica*, vol. 38, no. 9, pp. 1425–1432, 2002.

[35] H. S. Wilf, *Generatingfunctionology*. New York, NY: Academic Press, second ed., 1994.

[36] I. Kaminer, A. Pascoal, E. Hallberg, and C. Silvestre, "Trajectory tracking for autonomous vehicles: an integrated approach to guidance and control," *AIAA Journal of Guidance, Control, and Dynamics*, vol. 21, no. 1, pp. 29–38, 1998.

[37] R. Abraham, J. E. Marsden, and T. S. Ratiu, *Manifolds, Tensor Analysis, and Applications*, vol. 75 of *AMS*. New York, NY: Springer Verlag, second ed., 1988.

[38] C.-T. Chen, *Linear System Theory and Design*. New York, NY: Holt, Rinehart, and Winston, 1984.

[39] W. Press, W. Vetterling, S. Teukolsky, and B. Flannery, *Numerical Recipes in C*. New York, NY: Cambridge University Press, 1992.

[40] J. E. Hauser, S. S. Sastry, and G. Meyer, "Nonlinear control design for slightly nonminimum phase systems: application to V/STOL aircraft," *Automatica*, vol. 28, no. 4, pp. 665–679, 1992.

[41] P. Martin, S. Devasia, and B. Paden, "A different look at output tracking: Control of a VTOL aircraft," *Automatica*, vol. 32, no. 1, pp. 101–107, 1996.

[42] R. M. Murray, M. Rathinam, and W. Sluis, "Differential flatness of mechanical control systems: A catalog of prototype systems," in *ASME International Mechanical Engineering Congress and Exposition*, (San Francisco, CA), Nov. 1995.

[43] B. Char *et al.*, *Maple V Library Reference Manual*. Springer Verlag, 1991.

[44] M. T. Heath, *Scientific Computing: an Introductory Survey*. New York: McGraw-Hill, 2 ed., 2002.

Appendices

I. Solution of polynomial system as series expansion

We prove here Lemma II.1. *Proof:* The proof is a direct extension of the treatment in [34]. We present here only the convergence proof as the derivation of the formal expansion is straightforward. We start with the bounds

$$\|x_1\|_{\mathcal{L}_\infty} \leq \|e^{At}x_0\|_{\mathcal{L}_\infty} + \|e^{At}\|_{\mathcal{L}_1}\|Bu\|_{\mathcal{L}_\infty} = (1/2)d_1$$
$$\|x_k\|_{\mathcal{L}_\infty} \leq \|e^{At}\|_{\mathcal{L}_1}\|f^{[2]}\|_{\mathcal{L}_\infty}\sum_{i=1}^{k-1}\|x_i\|_{\mathcal{L}_\infty}\|x_{k-i}\|_{\mathcal{L}_\infty} = (1/2)d_2\sum_{i=1}^{k-1}\|x_i\|_{\mathcal{L}_\infty}\|x_{k-i}\|_{\mathcal{L}_\infty}.$$

Provided $d_1 d_2 \leq 1$, and assuming $\|x_k\|_{\mathcal{L}_\infty} \leq c_k d_1^k d_2^{k-1}$,

$$\|x_{k+1}\|_{\mathcal{L}_\infty} \leq (1/2)d_2\sum_{i=1}^{k}\|x_i\|_{\mathcal{L}_\infty}\|x_{k+1-i}\|_{\mathcal{L}_\infty} \leq (1/2)d_2\sum_{i=1}^{k}c_i c_{k+1-i}d_1^{k+1}d_2^{k-1} = c_{k+1}d_1^{k+1}d_2^k$$
$$\|x - \sum_{k=1}^{K}x_k\|_{\mathcal{L}_\infty} \leq \sum_{k=K+1}^{+\infty}\|x_k\|_{\mathcal{L}_\infty} \leq \sum_{k=K+1}^{+\infty}c_k d_1^k d_2^{k-1} = \frac{1}{d_2}\sum_{k=K+1}^{+\infty}c_k d_1^k d_2^k$$
$$\leq \frac{1}{d_2}\,\mathrm{Remainder}_K(\mathcal{C})(d_1 d_2).$$

Thus, Lemma II.1 is true by induction. ■

II. Convergence of iterative algorithm

We start with some preliminary results. Let $\mathrm{Symm}(\cdot)$ be the symmetrization operator [22] defined as

$$\mathrm{Symm}(F)(y_1, y_2, \ldots, y_k) = \frac{1}{k!}\sum_{\alpha_1,\ldots,\alpha_k\in\{0,1\}}(-1)^{(k-\sum_{i=1}^k\alpha_i)}F\left(\sum_{i=1}^k\alpha_i y_i, \ldots, \sum_{i=1}^k\alpha_i y_i\right). \tag{21}$$

Lemma II.1: Let F be a tensor, i.e., a multi-linear map, from k copies of $\mathbb{R}^n$ to $\mathbb{R}^n$. For all $y_1, y_2 \in \mathbb{R}^n$ we have

$$F(y_2, \ldots, y_2) - F(y_1, \ldots, y_1) = \sum_{j=0}^{k-1}\mathrm{Symm}(F)(y_2 - y_1, \underbrace{y_2, \ldots, y_2}_{k-1-j \text{ times}}, \underbrace{y_1, \ldots, y_1}_{j \text{ times}}). \tag{22}$$

Proof: Consider the following chain of equalities

$$\sum_{j=0}^{k-1}\mathrm{Symm}(F)(y_2-y_1,\underbrace{y_2,\dots,y_2}_{k-1-j\ \text{times}},\underbrace{y_1,\dots,y_1}_{j\ \text{times}})=$$

$$\sum_{j=0}^{k-1}\left(\mathrm{Symm}(F)(\underbrace{y_2,\dots,y_2}_{k-j\ \text{times}},\underbrace{y_1,\dots,y_1}_{j\ \text{times}})-\mathrm{Symm}(F)(\underbrace{y_2,\dots,y_2}_{k-1-j\ \text{times}},\underbrace{y_1,\dots,y_1}_{j+1\ \text{times}})\right)$$

$$=\sum_{j=0}^{k-1}\mathrm{Symm}(F)(\underbrace{y_2,\dots,y_2}_{k-j\ \text{times}},\underbrace{y_1,\dots,y_1}_{j\ \text{times}})-\sum_{i=1}^{k}\mathrm{Symm}(F)(\underbrace{y_2,\dots,y_2}_{k-i\ \text{times}},\underbrace{y_1,\dots,y_1}_{i\ \text{times}})$$

$$=\mathrm{Symm}(F)(\underbrace{y_2,\dots,y_2}_{k-j\ \text{times}},\underbrace{y_1,\dots,y_1}_{j\ \text{times}})\Big|_{j=0}-\mathrm{Symm}(F)(\underbrace{y_2,\dots,y_2}_{k-i\ \text{times}},\underbrace{y_1,\dots,y_1}_{i\ \text{times}})\Big|_{i=k}$$

$$=\mathrm{Symm}(F)(\underbrace{y_2,\dots,y_2}_{k\ \text{times}})-\mathrm{Symm}(F)(\underbrace{y_1,\dots,y_1}_{k\ \text{times}})=F(\underbrace{y_2,\dots,y_2}_{k\ \text{times}})-F(\underbrace{y_1,\dots,y_1}_{k\ \text{times}}).$$

■

Lemma II.2: For $\eta\in[0,1]$, the remainder of the Catalan function $\mathcal{C}(\eta)=1-\sqrt{1-\eta}$ satisfies

$$\mathrm{Remainder}_1\left(1-\sqrt{1-\eta}\right)=(1-\sqrt{1-\eta})-\frac{\eta}{2}\leq\frac{\eta^2}{2}.$$

Proof: Let $\eta\in[0,1]$. The following chain of inequalities holds

$$-\frac{3}{4}+\frac{1}{2}\eta+\frac{1}{4}\eta^2<0\quad\Rightarrow\quad-\frac{3}{4}\eta^2+\frac{1}{2}\eta^3+\frac{1}{4}\eta^4+(1-\eta)\leq(1-\eta)$$

$$\Rightarrow\quad\left(1-\frac{1}{2}\eta-\frac{1}{2}\eta^2\right)^2\leq(1-\eta)\quad\Rightarrow\quad\left(1-\frac{1}{2}\eta-\frac{1}{2}\eta^2\right)\leq\sqrt{1-\eta}\quad\Rightarrow\quad1-\sqrt{1-\eta}-\frac{1}{2}\eta\leq\frac{1}{2}\eta^2.$$

■

We are finally ready to prove Theorem IV.9, that we restate for convenience.

Theorem II.3: Let $z=2\|f_1^p\|_\infty D_1D_2\|x_{\text{target}}\|_\infty$. If

$$z<\min\left\{\frac{1}{D_1\|f_1^p\|_\infty},1-\frac{(D_1\|f_1^p\|_\infty)^2}{(1+D_1\|f_1^p\|_\infty)^2}\right\},$$

there exists a unique χ^* belonging to the set S and satisfying $\chi^*=\mathcal{M}(\chi^*)$. Furthermore, the unique solution can be computed by iterating the map $\mathcal{M}$ starting from any initial condition in S.

Proof: We prove the theorem in three steps. We show first that the series converges for any input in S, then that S is invariant under the map $\mathcal{M}$, and finally that $\mathcal{M}$ is a contraction over S.

First, note that $y=f_1^p\chi$ implies $\|y\|_\infty\leq\|f_1^p\|_\infty\|\chi\|_\infty$, and $\chi\in S$ implies $\|\chi\|_\infty\leq2\|x_{\text{target}}\|_\infty$. Hence we compose the bounds to obtain $D_1D_2\|y\|_\infty\leq D_1D_2\|f_1^p\|_\infty\|\chi\|_\infty\leq D_1D_2\|f_1^p\|_\infty2\|x_{\text{target}}\|_\infty=z<1-\frac{(D_1\|f_1^p\|_\infty)^2}{(1+D_1\|f_1^p\|_\infty)^2}\leq1$, which guarantees series convergence according to Lemma IV.1 and establishes that the contraction bounds are more conservative than the series bounds.

Second, we show that if $\chi \in S$, then $\mathcal{M}(\chi)$ also belongs to S, i.e., $\|\mathcal{M}(\chi) - x_{\text{target}}\|_\infty < \|x_{\text{target}}\|_\infty$. We compute

$$\begin{aligned}\|\mathcal{M}(\chi) - x_{\text{target}}\|_\infty &= \|\sum_{k=2}^{+\infty} f_k(f_1^p\chi, \ldots, f_1^p\chi)\|_\infty = \|f(f_1^p\chi) - f_1(f_1^p\chi)\|_\infty \\ &= \frac{1}{D_2}\,\text{Remainder}_1(\mathcal{C})(D_1D_2\|f_1^p\chi\|_\infty).\end{aligned}$$

From the bound in Lemma II.2

$$\begin{aligned}&= \frac{1}{D_2}\,\text{Remainder}_1(\mathcal{C})(D_1D_2\|f_1^p\chi\|_\infty) \leq \frac{(D_1D_2\|f_1^p\chi\|_\infty)^2}{2D_2} \\ &\leq \frac{(2D_1D_2\|f_1^p\|_\infty\|x_{\text{target}}\|_\infty)^2}{2D_2} = \frac{z^2}{2D_2}.\end{aligned}$$

From the second bound on z we have

$$\|\mathcal{M}(\chi) - x_{\text{target}}\|_\infty \leq \frac{z^2}{2D_2} = z(D_1\|f_1^p\|_\infty\|x_{\text{target}}\|_\infty) \leq \|x_{\text{target}}\|_\infty.$$

Finally, we show that $\|\mathcal{M}(\chi_2) - \mathcal{M}(\chi_1)\|_\infty \leq \rho\|\chi_2 - \chi_1\|_\infty$, where $0 \leq \rho < 1$. Applying the equality (22) from Lemma II.1:

$$\begin{aligned}\|\mathcal{M}(\chi_2) - \mathcal{M}(\chi_1)\|_\infty &= \|\sum_{k=2}^{+\infty}\sum_{j=0}^{k-1}\text{Symm}(f_k)(f_1^p(\chi_2-\chi_1), \underbrace{f_1^p\chi_2, \ldots, f_1^p\chi_2}_{k-1-j \text{ times}}, \underbrace{f_1^p\chi_1, \ldots, f_1^p\chi_1}_{j \text{ times}})\|_\infty \\ &\leq \sum_{k=2}^{+\infty}\sum_{j=0}^{k-1}\left(\|\text{Symm}(f_k)\|_\infty\|f_1^p\|_\infty^k\|\chi_2\|_\infty^{k-1-j}\|\chi_1\|_\infty^j\|\chi_2-\chi_1\|_\infty\right) \\ &\leq \|\chi_2-\chi_1\|_\infty\sum_{k=2}^{+\infty}\sum_{j=0}^{k-1}\left(\|\text{Symm}(f_k)\|_\infty\|f_1^p\|_\infty^k 2^{k-1}\|x_{\text{target}}\|_\infty^{k-1}\right) \\ &\leq \|\chi_2-\chi_1\|_\infty\sum_{k=2}^{+\infty}\left(k2^{k-1}\|\text{Symm}(f_k)\|_\infty\|f_1^p\|_\infty^k\|x_{\text{target}}\|_\infty^{k-1}\right). \qquad (23)\end{aligned}$$

We now upper bound $\|\text{Symm}(f_k)\|_\infty$ for $k > 1$

$$\|\text{Symm}(f_k)\|_\infty \leq (k!)^{-1}2^k\|f_k\|_\infty \leq 2^{1-k}2^k\|f_k\|_\infty = 2\|f_k\|_\infty.$$

Plugging the bound on f_k from Lemma IV.1 into equation (23), we obtain

$$\begin{aligned}\|\mathcal{M}(\chi_2) - \mathcal{M}(\chi_1)\|_\infty &\leq \|\chi_2-\chi_1\|_\infty\sum_{k=2}^{+\infty}\left(k2^k c_k D_1^k D_2^{k-1}\|f_1^p\|_\infty^k\|x_{\text{target}}\|_\infty^{k-1}\right) \\ &\leq \|\chi_2-\chi_1\|_\infty 2D_1\|f_1^p\|_\infty^k\sum_{k=2}^{+\infty}\left(kc_k z^{k-1}\right).\end{aligned}$$

The power series $\left(\sum_{k=1}^{+\infty} k a_k z^{k-1}\right)$ is the derivative of the generating function $\mathcal{C}$ (note the initial index), and can be shown to be convergent for $z < 1$. Using these facts, we can write

$$\|\mathcal{M}(\chi_2) - \mathcal{M}(\chi_1)\|_\infty \leq \|\chi_2 - \chi_1\|_\infty \, D_1 \|f_1^p\|_\infty^k \left(\frac{1}{\sqrt{1-z}} - 1\right) = \rho \|\chi_2 - \chi_1\|_\infty,$$

where we set $\rho = D_1 \|f_1^p\|_\infty^k \left(\frac{1}{\sqrt{1-z}} - 1\right)$. A few algebraic equalities based on last bound on z prove the bound $\rho < 1$. In summary, the map $\mathcal{M}$ is well-defined and is a contraction over the set S. The statement in the theorem follows from an application of the contraction mapping theorem. ■

III. Convergence of power series inversion

We start with some useful facts about a series.

Lemma III.1: Let $\beta \in \mathbb{R}_+$, consider the series of positive numbers

$$a_1 = 1, \qquad a_k = \beta \sum_{m=2}^{k} \sum_{\substack{i_1+\cdots+i_m=k \\ i_1,\cdots,i_m<k}} a_{i_1} \cdots a_{i_m}\,, \tag{24}$$

and define its generating function $h(\eta) = \sum_{k=1}^{+\infty} a_k \eta^k$. The following results hold:

(i) $h(\eta) = (1 + \eta - \sqrt{1 - 2(1+2\beta)\eta + \eta^2})/(2\beta + 2)$,

(ii) the function h is defined real, or in other words, the series $\sum_{k=1}^{+\infty} a_k \eta^k$ converges absolutely, provided $0 \leq \eta \leq (4(\beta+1))^{-1}$, and

(iii) the series $c_1 = \delta$, $c_k = \beta \sum_{m=2}^{k} \sum_{\substack{i_1+\cdots+i_m=k \\ i_1,\cdots,i_m<k}} \alpha^{m-1} c_{i_1} \cdots c_{i_m}$, can be bounded as $c_k \leq \delta^k \alpha^{k-1} a_k$.

We refer to [34] for the proof of most results in the lemma. Next, we prove Theorem IV.10. *Proof:* We start by showing $\Lambda_2 \leq \Lambda_1$. First, we have

$$\Lambda_2 = \frac{1}{4(D_1\|f_1^p\|_\infty + 1)\|f_1^p\|_\infty D_1 D_2} < \frac{1}{2\|f_1^p\|_\infty D_1 D_2}\left(\frac{1}{(D_1\|f_1^p\|_\infty + 1)}\right) \leq \frac{1}{2\|f_1^p\|_\infty D_1 D_2}\left(\frac{1}{D_1\|f_1^p\|_\infty}\right)$$

and furthermore

$$\begin{aligned}\Lambda_2 &< \frac{1}{2\|f_1^p\|_\infty D_1 D_2}\left(\frac{1}{(D_1\|f_1^p\|_\infty + 1)}\right) < \frac{1}{2\|f_1^p\|_\infty D_1 D_2}\left(\frac{1}{(D_1\|f_1^p\|_\infty + 1)}\right)\left(1 + \frac{D_1\|f_1^p\|_\infty}{(D_1\|f_1^p\|_\infty + 1)}\right) \\ &< \frac{1}{2\|f_1^p\|_\infty D_1 D_2}\left(\frac{1 + 2D_1\|f_1^p\|_\infty}{(1 + D_1\|f_1^p\|_\infty)^2}\right) < \frac{1}{2\|f_1^p\|_\infty D_1 D_2}\left(1 - \frac{(D_1\|f_1^p\|_\infty)^2}{(1 + D_1\|f_1^p\|_\infty)^2}\right).\end{aligned}$$

As seen in the proof of Theorem IV.9, when $\|x_{\text{target}}\|_\infty < \Lambda_1$, f is analytic (i.e., its series converges). Knowing this, we prove that the series defining the inverse function g in equation (18) converges uniformly in a neighborhood of x_{target}. From the Theorem IV.10, one can see that

$$\|g_1\|_\infty = \|f_1^{-1}\|_\infty\,, \qquad \|g_k\|_\infty \leq \|g_1\|_\infty \sum_{m=2}^{k} \sum_{\substack{i_1+\cdots+i_m=k \\ i_1,\cdots,i_m<k}} \|f_m\|_\infty \|g_{i_1}\|_\infty \cdots \|g_{i_m}\|_\infty.$$

Plugging in the bound on $\|f_m\|_\infty$ from Lemma IV.1, we have

$$\begin{aligned}\|g_k\|_\infty &\leq \|g_1\|_\infty \sum_{m=2}^{k} \sum_{\substack{i_1+\cdots+i_m=k \\ i_1,\cdots,i_m<k}} \left(c_m D_1^m D_2^{m-1}\right) \|g_{i_1}\|_\infty \cdots \|g_{i_m}\|_\infty \\ &\leq \left(D_1 \|f_1^{-1}\|_\infty\right) \sum_{m=2}^{k} \sum_{\substack{i_1+\cdots+i_m=k \\ i_1,\cdots,i_m<k}} (D_1 D_2)^{m-1} \|g_{i_1}\|_\infty \cdots \|g_{i_m}\|_\infty ,\end{aligned}$$

where we used the bound $c_k \leq 1$, for all $k > 1$. Let $\beta = D_1\|f_1^{-1}\|_\infty$, define the series $\{a_k \in \mathbb{R}, k \in \mathbb{N}\}$ as in equation (24), and following induction from the last statement above:

$$\|g_k\|_\infty \leq \|f_1^{-1}\|_\infty^k (D_1 D_2)^{k-1} a_k.$$

In summary, we have

$$\begin{aligned}\|g(x_{\text{target}})\|_\infty &\leq \| \sum_{k=1}^{+\infty} g_k(x_{\text{target}}, \ldots, x_{\text{target}})\|_\infty \\ &\leq \frac{1}{D_1 D_2} \sum_{k=1}^{+\infty} a_k \left(\|f_1^{-1}\|_\infty D_1 D_2 \|x_{\text{target}}\|_\infty\right)^k ,\end{aligned}$$

and, by Lemma III.1, convergence is ensured provided $4(D_1\|f_1^{-1}\|_\infty + 1)\|f_1^{-1}\|_\infty D_1 D_2 \|x_{\text{target}}\|_\infty \leq 1$.

Next, we prove that g is the inverse of f. The following proof is borrowed from [23] and we report it here for completeness. Evaluating the following expression,

$$\begin{aligned}f_1^{-1}\left(f\left(g(x)\right)\right) - g(x) &= f_1^{-1}\left(f - f_1\right)\left(g(x)\right) = f_1^{-1} \sum_{k=2}^{+\infty} f_k \left(\sum_{i_1=1}^{+\infty} g_{i_1}, \cdots \sum_{i_k=1}^{+\infty} g_{i_k} \right) \\ &= f_1^{-1} \sum_{k=2}^{+\infty} \sum_{i_1,\cdots i_k=1}^{+\infty} f_k\left(g_{i_1}, \cdots g_{i_k}\right) = \sum_{k=2}^{+\infty} f_1^{-1} \sum_{\substack{i_1+\cdots+i_m=k \\ i_1,\cdots,i_m<k}} f_m\left(g_{i_1}, \cdots g_{i_m}\right) \\ &= \sum_{k=2}^{+\infty} g_k = f_1^{-1}(x) - g(x).\end{aligned}$$

Therefore, $f_1^{-1}\left(f\left(g(x)\right)\right) = f_1^{-1}(x)$, and $\left(f\left(g(x)\right)\right) = x$. ■