Dynamic Optimization with Convergence Guarantees

Martin P. Neuenhofen and Eric C. Kerrigan

Abstract—We present a novel direct transcription method to solve optimization problems subject to non-linear differential and inequality constraints. In order to provide numerical convergence guarantees, it is sufficient for the functions that define the problem to satisfy boundedness and Lipschitz conditions. Our assumptions are the most general to date; we do not require uniqueness, differentiability or constraint qualifications to hold and we avoid the use of Lagrange multipliers. Our approach differs fundamentally from state-of-the-art methods based on collocation. We follow a least-squares approach to finding approximate solutions to the differential equations. The objective is augmented with the integral of a quadratic penalty on the differential equation residual and a logarithmic barrier for the inequality constraints, as well as a quadratic penalty on the point constraint residual. The resulting unconstrained infinite-dimensional optimization problem is discretized using finite elements, while integrals are replaced by quadrature approximations if they cannot be evaluated analytically. Order of convergence results are derived, even if components of solutions are discontinuous.

Index Terms—Dynamic optimization; infinite-dimensional optimization; optimal control; predictive control; moving horizon estimation; state estimation; parameter estimation; process optimization; trajectory optimization; finite element method; high-order methods; ordinary differential equations; differential-algebraic equations.

I. Introduction

A. An Important Class of Dynamic Optimization Problems

Many optimal control, estimation and system design problems can be written as a dynamic optimization problem in the form

$$\text{Find } (y^\star, z^\star) \in \underset{(y,z) \in \mathcal{X}}{\arg \min} \int_\Omega f(\dot{y}(t), y(t), z(t), t) \mathrm{d}t \quad \text{(DOPa)}$$

subject to

$$b\left(y(t_1), y(t_2), \ldots, y(t_M)\right) = 0, \quad \text{(DOPb)}$$
$$c\left(\dot{y}(t), y(t), z(t), t\right) = 0 \text{ f.a.e. } t \in \Omega, \quad \text{(DOPc)}$$
$$z(t) \geq 0 \text{ f.a.e. } t \in \Omega, \quad \text{(DOPd)}$$

where the open bounded interval $\Omega := (t_0, t_E) \subsetneq \mathbb{R}$, $\mathcal{X}$ is an appropriately-defined Hilbert space such that y is continuous and "f.a.e." means "for almost every" in the Lebesgue sense; detailed definitions and assumptions are

M.P. Neuenhofen is with the Department of Computer Science, University of British Columbia, Vancouver BC, e-mail: see http://www.MartinNeuenhofen.de

E.C. Kerrigan is with the Department of Electrical & Electronic Engineering and the Department of Aeronautics, Imperial College London, SW7 2AZ London, UK, e-mail: e.kerrigan@imperial.ac.uk (see http://www.imperial.ac.uk/people/e.kerrigan).

given in Section I-D. We note that the form in (DOP) is quite general. As shown in Appendix A, problems in the popular Bolza form with general inequalities and those with a free initial-time t_0 or end-time t_E, can be converted into (DOP); in turn, many problems from control and state estimation can be stated in Bolza form [1].

Dynamic equations and path constraints are included via the differential-algebraic equations (DAE) in (DOPc) and the inequalities (DOPd). The point constraints (DOPb) enforce boundary constraints, such as initial or final values on the state y, or can be used to include values obtained by measurements at given points. Constant parameters that are to be determined, or unknown parameters that are to be estimated, can also be included in y. The free variable z can include manipulated control inputs to be determined or unknown external signals, such as measurement or process noise.

Problem (DOP) is an infinite-dimensional optimization problem, because the optimization is over function spaces subject to an uncountable set of constraints. It is very hard or impossible to compute an analytic solution, in general. Hence, for many problems one has to resort to a numerical method to compute an approximate solution. When doing so, it is important to eliminate whether features of the solution have arisen from physical principles or numerical errors. The need for a numerical method, which has a rigorous proof that the approximate solution convergences to the actual solution, is therefore essential when solving practical problems.

The state-of-the-art for the numerical solution of (DOP) is to discretize via direct collocation with finite elements [1]–[5]. However, as will be shown in Section V via numerical examples, collocation methods can fail to converge if care is not taken.

Recall that explicit Runge-Kutta methods are unsuitable for stiff problems and that most popular implicit methods for solving differential equations, e.g. variants of Gauss schemes, can be interpreted as collocation methods. Collocation methods include certain classes of implicit Runge-Kutta methods, pseudospectral methods, as well as Adams and backward differentiation formula (BDF) methods [1], [3], [5], [6].

There is a scarcity of rigorous proofs that high-order collocation schemes for dynamic optimization methods converge to a feasible or optimal solution as the discretization is refined. The assumptions in the literature are often highly technical, difficult to enforce or not very general.

B. Contributions

This paper presents a novel direct transcription method for solving (DOP), together with a proof of convergence,

which has the most general assumptions to date. As detailed in Section I-D, we will only require existence of a solution to (DOP) and mild assumptions on the boundedness and Lipschitz continuity of f, c, b. In contrast to existing convergence results:

- The solution $(y^\star, z^\star)$ does not need to be unique.
- f, c, b need not be differentiable anywhere.
- We do not require the satisfaction of a constraint qualification for the discretized, finite-dimensional optimization problem, such as the Linear Independence Constraint Qualification (LICQ), Mangasarian-Fromovitz Constraint Qualification (MFCQ) or Second Order Sufficient Conditions (SOSC).
- We do not assume uniqueness or global smoothness of states or co-states/adjoints.
- We remove the assumptions in [7] on local uniqueness.

To discuss the main idea behind our method, recall that for a practical method one wants to find a weak numerical solution $x_h^\epsilon \in \mathcal{X}$ that solves (DOP) in a *tolerance-accurate* sense. To make this precise, define the objective functional

$$F(x) := \int_\Omega f\left(\dot{y}(t), y(t), z(t), t\right) \mathrm{d}t$$

and a measure of feasibility for elements $x := (y, z) \in \mathcal{X}$:

$$r(x) := \int_\Omega \|c\left(\dot{y}(t), y(t), z(t), t\right)\|_2^2 \, \mathrm{d}t \\ + \|b\left(y(t_1), y(t_2), \ldots, y(t_M)\right)\|_2^2,$$

where $\| \cdot \|_2$ denotes the usual 2-norm of a vector. The optimality gap g_{opt} and feasibility residual r_{feas} of x_h^ϵ are defined as

$$g_{\mathrm{opt}} := \max\{F(x_h^\epsilon) - F(x^\star), 0\}, \qquad r_{\mathrm{feas}} := r(x_h^\epsilon).$$

The goal of our numerical method is the construction of an x_h^ϵ such that this gap and residual can be driven below an arbitrary, strictly positive tolerance. In addition, our numerical solution will satisfy $z_h^\epsilon(t) > 0 \ \forall t \in \overline{\Omega}$.

The method starts in the same way as collocation methods by defining a Finite Element space. The approximate solution is constrained to a finite dimensional subspace $\mathcal{X}_{h,p} \subset \mathcal{X}$, namely piecewise polynomial functions on a mesh, where h denotes the size of the largest interval in the mesh and p is the maximum degree of the polynomials in each mesh interval. However, beyond using finite elements, our penalty approach is entirely different from collocation.

Collocation methods replace the uncountable set of constraints (DOPc) by a finite set of constraints, called collocation constraints, namely

$$c\left(\dot{y}(t), y(t), z(t), t\right) = 0, \ \forall t \in \mathbb{T}_{h,p},$$

where $\mathbb{T}_{h,p}$ is a finite set of points in $[t_0, t_E]$, called collocation points. Since the differential equation is satisfied at only a finite number of points and $\mathcal{X}_{h,p}$ is a strict subset of $\mathcal{X}$, the feasibility residual r_{feas} is non-zero, in general. If the residual is too large, then the dimension of $\mathcal{X}_{h,p}$ is usually increased in correspondence with choosing a set $\mathbb{T}_{h,p}$ with more elements. It is well-known that if care

is not taken with the choice of $\mathbb{T}_{h,p}$ and $\mathcal{X}_{h,p}$, then the feasibility residual will not converge, e.g. due to Runge's phenomenon, or a solution might not exist, e.g. when there are more constraints than degrees of freedom. Essentially, problems arise in collocation methods because they do not explicitly take into account what happens in-between collocation points and the feasibility residual is usually calculated a posteriori. Furthermore, most convergence results typically assume that c and/or the solution is sufficiently smooth. Unfortunately, this cannot be guaranteed for inequality-constrained dynamic optimization problems, where $y^\star$ may be continuous but with non-smooth derivative, and $z^\star$ may be discontinuous, hence the shortage of results on this topic. Finally, note that the set of collocation points $\mathbb{T}_{h,p}$ is, in general, not a function of the problem data or the solution, hence the collocation points are not optimal in terms of minimizing the feasibility residual r_{feas}.

Our method differs fundamentally from collocation methods by explicitly considering the violation of constraints in the whole of Ω when formulating the optimization problem. This is possible due to the observation that one can replace the uncountable set of constraints (DOPc) with the finite set of constraints

$$\int_T \|c\left(\dot{y}(t), y(t), z(t), t\right)\|_2^2 \mathrm{d}t \leq \varepsilon_T, \ \forall T \in \mathcal{T}_h. \quad \text{(DOPc')}$$

Here $\mathcal{T}_h$ is a finite set of disjoint intervals such that $\bigcup_{T \in \mathcal{T}_h} \overline{T} = [t_0, t_E]$. Clearly, (DOPc') is satisfied with $\varepsilon_T = 0$ if and only if (DOPc) is satisfied; however, since $\mathcal{X}_{h,p}$ is finite dimensional, usually the best that one can do is to satisfy the above constraint with $\varepsilon_T > 0$. The integral can either be obtained analytically or approximated to high accuracy with numerical quadrature.

The approach we therefore adopt is to augment the objective functional with a weighted version of the integrals in (DOPc') and a quadratic penalty on the equality constraints (DOPb), namely a scaled version of r as defined above. We will show in Section III that the feasibility residual, hence all ε_T, can be driven arbitrary close to zero as the number of elements in $\mathcal{T}_h$ increases. The conditions that we provide on $\mathcal{T}_h$ are simpler than the conditions on $\mathbb{T}_{h,p}$ in the collocation literature; note that $\mathcal{T}_h$ will not be a function of the degree p of the polynomials.

The inequality constraints (DOPd) also need careful consideration. We will treat these by augmenting the objective functional with the *integral* of a logarithmic barrier, which can be evaluated analytically if desired, to ensure that constraints are satisfied over the whole interval Ω. In contrast, collocation methods usually enforce the inequality constraints at only a finite subset of Ω.

This paper not only provides a method and conditions under which convergence can be guaranteed, but also provides *order* of convergence results. We are able to do this because we first formulate a related infinite-dimensional, unconstrained problem. We show that solutions of this unconstrained problem converge to solutions of (DOP). We then show that solutions of the discretized, unconstrained

problem converge to solutions of (DOP). In practice, the solutions of the infinite-dimensional unconstrained problem are usually piecewise smooth and, as a consequence, one can obtain high-order convergence guarantees.

In the special case when $f = 0$ and the positivity constraints (DOPd) are removed, then (DOP) reduces to finding a solution of the differential equations (DOPc) subject to (DOPb). Our method then becomes that of solving a sequence of finite-dimensional, unconstrained least-squares problems $\min_{x \in \mathcal{X}_{h,p}} r(x)$, where the mesh is refined until the feasibility residual r_{feas} is below the required tolerance. Least-squares methods for solving linear and nonlinear ordinary differential equations, with high-order convergence guarantees, have been available for some time [8], [9]. This paper can be interpreted as a generalization of this approach to solving dynamic optimization problems.

C. Literature Review

The literature on convergence of numerical methods for the solution of ordinary differential equations (ODE) and DAEs is substantial. This is not the case for dynamic optimization problems.

A convergence proof for a discretization based on the explicit Euler method is given in [10]. The result makes the following strong assumptions: (i) functions defining the problem must be locally differentiable with Lipschitz continuous derivatives; (ii) there must be a local solution where the trajectories of the state and free variables are continuously differentiable and continuous, respectively; (iii) a homogeneity condition on active constraints; (iv) surjectivity of linearized equality constraints; and (v) a coercivity assumption.

These conditions are sophisticated, difficult to understand, and very hard to verify and ensure by construction. These conditions ensure that the first order optimality conditions of the infinite-dimensional problem result in a unique and stable solution for the optimality conditions of the collocation method. This is why convergence proofs for other collocation schemes make similar assumptions. Furthermore, since Euler's method converges only to first order, convergence results based on Euler's method are of limited practical use compared to results for higher-order methods, if such results are available.

In [11] the authors propose an ℓ_1-penalty method for a high-order collocation method and demonstrate the stabilization effect of this approach for a numerical test problem, namely the Aly-Chan problem [12] in Mayer form with path constraints. The result is experimental. The discussion in [11] assumes similar conditions as in [13], where they present a convergence result for problems without path constraints and do not use the ℓ_1-penalty method from [11]. The result in [13] relies on a number of strong assumptions: (i) functions defining the problem must be sufficiently smooth; (ii) the state and co-state trajectories must be sufficiently smooth; (iii) the nonlinear program arising from the discretization must satisfy LICQ and SOSC.

Another proof of high-order convergence for a direct collocation method is given in [14]. This proof considers the Bolza form with constraints. The authors show convergence of their scheme under the following assumptions: (i) the solution must be locally unique; (ii) the states are assumed to have a strong first derivative; (iii) the state and co-state trajectories must have two square integrable derivatives; (iv) the Hamiltonian is assumed to satisfy a local strong convexity property; (v) the objective and the ODE function are assumed to be twice Lipschitz differentiable.

A convergence result for a pseudospectral method for the control of constrained feedback linearizable systems is given in [15]. In order to provide a proof that does not require dualization, the strong assumption is made that the derivatives of the interpolating polynomial for the state converges uniformly to a continuous function, which implies that the optimal input has to be continuous. This assumption was relaxed in [16] to allow for discontinuous optimal inputs, under the assumptions that (i) both the optimal state and input trajectories are piecewise differentiable, and (ii) the path constraints define a convex set.

The assumption on feedback linearizable systems in [15], [16] was relaxed in [17] to allow for more general nonlinear ordinary differential equations. However, the following strong assumptions are made: (i) the functions defining the problem are continuously differentiable, (ii) the gradients of the functions are Lipschitz continuous, (iii) the optimal state trajectory is continuously differentiable, which requires that the optimal control trajectory be continuous.

Direct multiple shooting [1], [3], [18] is an effective method for solving dynamic optimization problems. Convergence proofs for shooting methods for quadratic regulator problems with linear dynamics and linear inequality constraints are presented in [19]–[21], where the differential equation is solved using matrix exponentials. For more general nonlinear systems, depending on the discretization and integration method implemented, e.g. explicit/implicit Runge-Kutta or certain classes of collocation methods, one could interpret certain existing results as a proof of convergence for a direct multiple shooting.

A high-order convergence proof for explicit, fixed-step size Runge-Kutta methods is given in [22], which extends some of the results presented in [23] for the Euler method. The assumptions in [22] are that (i) the inputs are constrained to lie in a convex and compact norm-ball (ii) the state derivatives can be written as an explicit, continuously differentiable function of the state and input, and that (iii) the dynamic equation is time-invariant.

Convergence results for Runge-Kutta methods for unconstrained optimal control problems are available in [24], subject to the following assumptions: (i) the optimal states and their first and second derivatives are globally bounded, (ii) the optimal control is continuously differentiable, (iii) functions defining the problem are twice differentiable, and (iv) a local coercivity property of the linear-quadratic approximation in the local minimizer.

Indirect methods have also been widely studied for com-

puting the solution of (DOP) [1], [3]. In these, the calculation of variations is used to determine the optimality conditions for the optimal arcs. Indirect methods have been less successful in practice than direct methods. Firstly, for complicated problems it is difficult to determine optimality conditions in a correct way. Secondly, for singular-arc problems, the first-order optimality conditions are insufficient for determining the optimal solution. Thirdly, the optimality conditions can be ill-posed, for example when the co-state solution is non-unique. This always happens when path-constraints are collinear. Finally, the resulting optimality system, called a Hamiltonian boundary-value problem (HBVP), is difficult to solve numerically. Methods for treating these have issues with robustness and may easily fail when no accurate initial guess for the solution of the HBVP is given [25]. For these reasons, we do not consider a detailed review of indirect methods.

D. Notation and Assumptions

Let $-\infty < t_0 < t_E < \infty$ and the $M \in \mathbb{N}$ points $t_k \in \overline{\Omega}$, $\forall k \in \{1, 2, \ldots, M\}$. $\overline{\Omega}$ denotes the closure of a set Ω. The functions $f : \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \times \Omega \to \mathbb{R}$, $c : \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \times \Omega \to \mathbb{R}^{n_c}$, $b : \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \times \ldots \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_b}$. The function $y : \overline{\Omega} \to \mathbb{R}^{n_y}, t \mapsto y(t)$ and $z : \overline{\Omega} \to \mathbb{R}^{n_z}, t \mapsto z(t)$. Given an interval $\Omega \subset \mathbb{R}$, let $|\Omega| := \int_\Omega 1 \, \mathrm{d}t$. We use Big-$\mathcal{O}$ notation to analyze a function's behaviour close to zero, i.e. function $\phi(\xi) = \mathcal{O}(\gamma(\xi))$ if and only if $\exists C > 0$ and $\xi_0 > 0$ such that $|\phi(\xi)| \le C\gamma(\xi)$ when $0 < \xi < \xi_0$. The vector $\mathbf{1} := [1 \; \cdots \; 1]^\mathsf{T}$ with appropriate size.

For notational convenience, we define the function

$$x := (y, z) : \overline{\Omega} \to \mathbb{R}^{n_x},$$

where $n_x := n_y + n_z$. The solution space of x is the Hilbert space

$$\mathcal{X} := \left(H^1\left(\Omega\right) \right)^{n_y} \times \left(L^2\left(\Omega\right) \right)^{n_z}$$

with scalar product

$$\langle (y, z), (v, w) \rangle_\mathcal{X} := \sum_{j=1}^{n_y} \langle y_{[j]}, v_{[j]} \rangle_{H^1(\Omega)} + \sum_{j=1}^{n_z} \langle z_{[j]}, w_{[j]} \rangle_{L^2(\Omega)}$$

and induced norm $\| \cdot \|_\mathcal{X}$, where $\phi_{[j]}$ denotes the jth component of a function ϕ. The Sobolev space $H^1(\Omega)$ and Lebesgue space $L^2(\Omega)$ are defined as usual [26]. The weak derivative of y is denoted by $\dot{y} := \mathrm{d}y/\mathrm{d}t$.

Recall the embedding $H^1(\Omega) \hookrightarrow \mathcal{C}^0(\overline{\Omega})$, where $\mathcal{C}^0(\overline{\Omega})$ denotes the space of continuous functions over $\overline{\Omega}$ [26, Thm 5.4, part II, eqn 10]. Hence, by requiring that $y \in \left(H^1\left(\Omega\right) \right)^{n_y}$ it follows that y is continuous. In contrast, though $\dot{y}$ and z are in $L^2(\Omega)$, they may be discontinuous.

We make the following assumptions on (DOP):

(A.1) (DOP) has a feasible point.

(A.2) $\|f\|_1, \|c\|_1, \|b\|_1$ are globally bounded in terms of the essential supremum.

(A.3) f, c, b are globally Lipschitz continuous in all arguments, except t, with respect to $\| \cdot \|_1$.

(A.4) Two solutions $x^\star_\omega, x^\star_{\omega,\tau}$ related to $x^\star$, defined in Section II, are bounded in terms of $\|z\|_{L^\infty(\Omega)}$ and $\|x\|_\mathcal{X}$. Also, $\|x^\star\|_\mathcal{X}$ is bounded.

(A.5) The related solution $x^\star_{\omega,\tau}$ can be approximated to high order using piecewise polynomials; formalized in (4).

We discuss the assumptions. (A.1) is reasonable. (A.2)–(A.3) are mild and can be ensured by construction. To this end, f, c, b can be bounded below and/or above, if necessary, with minimum and maximum terms. Functions that are not Lipschitz continuous, e.g. the square-root, can be made so by replacing them with smoothed functions, e.g. via a mollifier. Smoothing is a common practice to ensure the derivatives used in a nonlinear optimization algorithm (e.g. IPOPT [27]) are globally well-defined. (A.4) can be ensured as shown in Section II-B Remark 1. (A.5) is rather mild, as discussed in Section III-D and illustrated in Appendix C.

The assumptions are not necessary but sufficient. Suppose that we have found a numerical solution. It is not of relevance to the numerical method whether the assumptions hold outside of an open neighborhood of this solution. However, the proofs below would become considerably more complicated with local assumptions, hence why we opted for global assumptions. As will be seen in Section V, convergence is obtained in numerical examples even if the assumptions on f, c, b hold only locally, whereas certain collocation methods fail to converge.

Furthermore, our assumptions are more general than those detailed in Section I-C, which can be technical and difficult to verify or enforce.

E. Outline

Section II introduces a reformulation of (DOP) as an unconstrained problem. Section III presents the Finite Element Method in order to formulate a finite-dimensional unconstrained optimization problem. The main result of the paper is Theorem 3, which shows that solutions of the finite-dimensional optimization problem converge to solutions of (DOP) with a guarantee on the order of convergence. Section IV briefly discusses how one could compute a solution using nonlinear programming (NLP) solvers. Section V presents numerical results which validate that our method converges for difficult problems, whereas certain collocation methods can fail in some cases. Conclusions are drawn in Section VI.

II. Reformulation as an Unconstrained Problem

The reformulation of (DOP) into an unconstrained problem is achieved in two steps. We first introduce penalties for the differential-algebraic and equality constraints and then add logarithmic barriers for the inequality constraints. The motivation is that convergence theory for the finite element method of Section III is simpler for unconstrained minimizers.

Before proceeding, we note that boundedness and Lipschitz-continuity of F and r follow from (A.2)–(A.3).

Lemma 1 (Boundedness and Lipschitz-continuity of F and r)**.** *Both F and r are bounded and Lipschitz continuous in x with respect to $\|\cdot\|_{\mathcal{X}}$. Furthermore, F and r are Lipschitz continuous in z with respect to the norm $\|\cdot\|_{L^1(\Omega)}$.*

Proof: Boundedness of F, r follows from (A.2).

Lipschitz continuity of r is not as straightforward. We will make use of the following trace theorem [28]: For an open interval $I \subseteq \Omega$ it holds that $\|u\|_{L^2(\partial I)} \leq K \cdot \|u\|_{H^1(I)}$ with a constant K independent of u. Assume $|u|$ attains its essential supremum on $\overline{\Omega}$ at $t = t^\star$. Choosing $I = (t^\star, t_E) \subset \Omega$, then $\|u\|_{L^\infty(\Omega)} = |u(t^\star)| \leq \|u\|_{L^2(\partial I)}$. Using this together with the above bound and $\|u\|_{H^1(I)} \leq \|u\|_{H^1(\Omega)}$ results in

$$\|u\|_{L^\infty(\Omega)} \leq K \cdot \|u\|_{H^1(\Omega)}. \tag{1}$$

Below, for a generic Lipschitz continuous function $g : \mathbb{R}^k \to \mathbb{R}^{n_g}$ with Lipschitz-constant L_g and $\|\cdot\|_1$-bound $|g|_{\max}$, we use the relation

$$
\begin{aligned}
&\big| \|g(\xi_2)\|_2^2 - \|g(\xi_1)\|_2^2 \big| \\
&= \big| \|g(\xi_2)\|_2 + \|g(\xi_1)\|_2 \big| \cdot \underbrace{\big| \|g(\xi_2)\|_2 - \|g(\xi_1)\|_2 \big|}_{\leq \|g(\xi_2) - g(\xi_1)\|_2} \\
&\leq n_g \cdot \big| \|g(\xi_2)\|_1 + \|g(\xi_1)\|_1 \big| \cdot \|g(\xi_2) - g(\xi_1)\|_1 \\
&\leq n_g \cdot 2 \cdot |g|_{\max} \cdot L_g \cdot \|\xi_2 - \xi_1\|_1 \,,
\end{aligned}
$$

where we used $|\alpha^2 - \beta^2| = |\alpha + \beta| \cdot |\alpha - \beta|$ in the first line and the triangular inequality in the second line. Using the above generic bound, we can show Lipschitz continuity of r:

$$
\begin{aligned}
|r(x_2) - r(x_1)| &\leq \int_\Omega \Big| \|c(\dot{y}_2(t), y_2(t), z_2(t), t)\|_2^2 \\
&\qquad\qquad - \|c(\dot{y}_1(t), y_1(t), z_1(t), t)\|_2^2 \Big| \mathrm{d}t \\
&\quad + \Big| \|b(y_2(t_1), \ldots, y_2(t_M))\|_2^2 - \|b(y_1(t_1), \ldots, y_1(t_M))\|_2^2 \Big| \\
&\leq \int_\Omega 2 \cdot n_c \cdot |c|_{\max} \cdot L_c \cdot \left\| \begin{pmatrix} \dot{y}_2(t) - \dot{y}_1(t) \\ y_2(t) - y_1(t) \\ z_2(t) - z_1(t) \end{pmatrix} \right\|_1 \mathrm{d}t \\
&\quad + 2 \cdot n_b \cdot |b|_{\max} \cdot L_b \cdot \underbrace{\left\| \begin{pmatrix} y_2(t_1) - y_1(t_1) \\ \vdots \\ y_2(t_M) - y_1(t_M) \end{pmatrix} \right\|_1}_{\leq M \cdot \|y_2 - y_1\|_{L^\infty(\Omega)}} \\
&\leq 2 \cdot n_c \cdot |c|_{\max} \cdot L_c \cdot \\
&\quad \big(\|\dot{y}_2 - \dot{y}_1\|_{L^1(\Omega)} + \|y_2 - y_1\|_{L^1(\Omega)} + \|z_2 - z_1\|_{L^1(\Omega)} \big) \\
&\quad + 2 \cdot n_b \cdot |b|_{\max} \cdot L_b \cdot M \cdot K \cdot \|y_2 - y_1\|_{H^1(\Omega)},
\end{aligned}
$$

where (1) has been used in the last line to bound $\|y_2 - y_1\|_{L^\infty(\Omega)}$.

If $y_2 = y_1$ then we see the result shows Lipschitz continuity of r with respect to $\|z\|_{L^1(\Omega)}$. Using

$$\|u\|_{L^1(\Omega)} \leq \sqrt{t_E - t_0} \cdot \|u\|_{L^2(\Omega)} \quad \forall u \in L^1(\Omega)$$

according to [26, Thm. 2.8, eqn. 8], and the definition of $\|\cdot\|_{\mathcal{X}}$, we arrive at

$$
\begin{aligned}
&\|\dot{y}_2 - \dot{y}_1\|_{L^1(\Omega)} + \|y_2 - y_1\|_{L^1(\Omega)} + \|z_2 - z_1\|_{L^1(\Omega)} \\
&\leq \sqrt{|\Omega|} \cdot \big(\|\dot{y}_2 - \dot{y}_1\|_{L^2(\Omega)} + \|y_2 - y_1\|_{L^2(\Omega)} + \|z_2 - z_1\|_{L^2(\Omega)} \big) \\
&\leq \sqrt{|\Omega|} \cdot \|x_2 - x_1\|_{\mathcal{X}},
\end{aligned}
$$

which shows Lipschitz continuity of r with respect to $\|x\|_{\mathcal{X}}$.

The proof for Lipschitz continuity of F follows from Lipschitz-continuity of f:

$$
\begin{aligned}
&|F(x_2) - F(x_1)| \\
&\leq \int_\Omega |f(\dot{y}_2(t), y_2(t), z_2(t)) - f(\dot{y}_1(t), y_1(t), z_1(t))| \, \mathrm{d}t \\
&\leq \int_\Omega L_f \cdot \left\| \begin{pmatrix} \dot{y}_2(t) - \dot{y}_1(t) \\ y_2(t) - y_1(t) \\ z_2(t) - z_1(t) \end{pmatrix} \right\|_1 \mathrm{d}t \leq L_f \cdot \left\| \begin{pmatrix} \dot{y}_2 - \dot{y}_1 \\ y_2 - y_1 \\ z_2 - z_1 \end{pmatrix} \right\|_{L^1(\Omega)}
\end{aligned}
$$

∎

We bound the Lipschitz constants (i.e., with respect to $\|x\|_{\mathcal{X}}$ and $\|z\|_{L^1(\Omega)}$) with $L_F \geq 2$ for F and with $L_r \geq 2$ for r.

A. Penalty Form

We introduce the *penalty problem*

Find $x_\omega^\star \in \underset{x \in \mathcal{X}}{\arg\min}\, F_\omega(x)$ s.t. $z(t) \geq 0$ f.a.e. $t \in \Omega$, (PP)

where

$$F_\omega(x) := F(x) + \frac{1}{2 \cdot \omega} \cdot r(x)$$

and the penalty parameter $\omega \in (0, 0.5]$. Note that F_ω is Lipschitz continuous with constant

$$L_\omega := L_F + \frac{1}{2 \cdot \omega} \cdot L_r.$$

We show that ε-optimal solutions of (PP) solve (DOP) in a tolerance-accurate way.

Proposition 1 (Penalty Solution)**.** *Let $\varepsilon \geq 0$. Consider an ε-optimal solution x_ω^ε to (PP), i.e.*

$$F_\omega(x_\omega^\varepsilon) \leq F_\omega(x_\omega^\star) + \varepsilon \text{ and } z_\omega^\varepsilon(t) \geq 0 \text{ f.a.e. } t \in \Omega\,.$$

If we define $C_r := 2 \cdot \operatorname{ess\,sup}_{x \in \mathcal{X}} |F(x)|$, then

$$
\begin{aligned}
F(x_\omega^\varepsilon) &\leq F(x^\star) + \varepsilon\,, \\
r(x_\omega^\varepsilon) &\leq 2 \cdot \omega \cdot (C_r + \varepsilon)\,.
\end{aligned}
$$

Proof: $x^\star$, $x_\omega^\star$, x_ω^ε are all feasible for (PP), but $x_\omega^\star$ is optimal and x_ω^ε is ε-optimal. Thus,

$$\underbrace{F_\omega(x_\omega^\varepsilon)}_{\geq F(x_\omega^\varepsilon)} \leq \underbrace{F_\omega(x^\star)}_{= F(x^\star)} + \varepsilon\,. \tag{2}$$

From this follows $F(x_\omega^\varepsilon) \leq F(x^\star) + \varepsilon$ because $r(x_\omega^\varepsilon) \geq 0$ and $r(x^\star) = 0$ by (A.1). To show the second proposition, subtract $F(x_\omega^\varepsilon)$ from (2). Then it follows that $1/(2 \cdot \omega) \cdot r(x_\omega^\varepsilon) \leq F(x^\star) - F(x_\omega^\varepsilon) + \varepsilon \leq C_r + \varepsilon$. Multiplication of this inequality with $2 \cdot \omega$ shows the result.

The fact that C_r is bounded follows from Lemma 1. ∎

This result implies that for an ε-optimal solution to (PP) the optimality gap to (DOP) is less than ε and that the feasibility residual can be made arbitrarily small by choosing the penalty parameter ω to be sufficiently small.

In Proposition 1 we used (A.2) which implies $|F|$ is bounded. In fact, F only needs to be bounded below. To show this, note that

$$r(x_\omega^\varepsilon) = 2 \cdot \omega \cdot \big(\underbrace{F_\omega(x_\omega^\varepsilon)}_{\leq F(x^\star)+\varepsilon} - \underbrace{F(x_\omega^\varepsilon)}_{\geq F_{lb}} \big).$$

Hence, $r(x_\omega^\varepsilon) \leq 2 \cdot \omega \cdot \big(\underbrace{F(x^\star) - F_{lb}}_{=\mathcal{O}(1)} + \varepsilon \big).$

B. Penalty-Barrier Form

We reformulate (PP) once more in order to remove the inequality constraints. We do so using logarithmic barriers. Consider the *penalty-barrier problem*

$$\text{Find } x_{\omega,\tau}^\star \in \arg\min_{x \in \mathcal{X}} F_{\omega,\tau}(x) := F_\omega(x) + \tau \cdot \Gamma(x), \quad \text{(PBP)}$$

where the barrier parameter $\tau \in (0,\omega]$ and

$$\Gamma(x) := -\sum_{j=1}^{n_z} \int_\Omega \log\big(z_{[j]}(t)\big) \, \mathrm{d}t.$$

We have introduced logarithmic barriers in order to keep $z_{\omega,\tau}^\star$ strictly positive. Recall that $L^2(\Omega)$ contains functions that have poles. It is therefore reasonable to ask whether these logarithmic barriers will ensure that the components of $z_{\omega,\tau}^\star$ are non-negative. This question motivates the following result.

Lemma 2 (Strict Interiorness).

$$z_{\omega,\tau}^\star(t) \geq \frac{\tau}{L_\omega} \cdot \mathbf{1} \; f.a.e. \; t \in \Omega.$$

Proof: We consider a worst-case example, where z is as close as possible to 0 in every component at almost every $t \in \Omega$. Since F_ω is Lipschitz continuous in z with respect to $\|\cdot\|_{L^1(\Omega)}$, this worst-case example would be obtained if $F_\omega(x) = L_\omega \cdot \|z\|_{L^1(\Omega)}$. Consequently, $F_{\omega,\tau}(x) = L_\omega \cdot \|z\|_{L^1(\Omega)} + \tau \cdot \Gamma(x)$. The minimizer of $F_{\omega,\tau}$ is $z_{[j]}(t) = \frac{\tau}{L_\omega}$ f.a.e. $t \in \Omega$, $\forall j \in \{1, 2, \ldots, n_z\}$. Hence, in general τ/L_ω is an essential lower bound. ∎

We will need the following algebraic result, which uses an arbitrary fixed number $0 < \zeta \ll 1$.

Lemma 3 (Order of the log Term).

$$|\tau \cdot \log(\tau/L_\omega)| = \mathcal{O}\left(\tau^{1-\zeta}\right).$$

Proof: We use $L_\omega = L_F + \frac{1}{2 \cdot \omega} \cdot L_r$, where $L_F \geq 2$, $L_r \geq 2$ and $0 < \tau \leq \omega \leq 0.5$. We get

$$|\tau \cdot \log(\tau/L_\omega)|$$
$$= \tau \cdot (|\log(\tau) - \log(L_\omega)|) \leq \tau \cdot (|\log(\tau)| + |\log(L_\omega)|)$$
$$= \tau \cdot \left(\left|\log\left(L_F + \frac{L_r}{2 \cdot \omega}\right)\right| + |\log(\tau)| \right)$$
$$\leq \tau \cdot \left(1 + |\log(L_F)| + \left|\log\left(\frac{L_r}{2 \cdot \omega}\right)\right| + |\log(\tau)| \right)$$
$$\leq \tau \cdot \Big(\underbrace{1 + |\log(L_F)| + |\log(L_r/2)|}_{=\mathcal{O}(1)} + \underbrace{|\log(\omega)|}_{\leq |\log(\tau)|} + |\log(\tau)| \Big)$$
$$= \mathcal{O}(\tau) + \mathcal{O}(\tau \cdot |\log(\tau)|).$$

In the third line above, we used the fact that for $\alpha, \beta \geq 2$, (where w.l.o.g. we assume $\alpha \geq \beta$) we have $\log(\alpha + \beta) \leq \log(2 \cdot \alpha) < \log(\alpha) + 1 < \log(\alpha) + 1 + \log(\beta)$.

The result follows because $\tau \cdot |\log(\tau)| = \mathcal{O}(\tau^{1-\zeta})$, as we show using L'Hôpital's rule:

$$\lim_{\tau \to 0} \frac{\tau \cdot \log(\tau)}{\tau^{1-\zeta}} = \lim_{\tau \to 0} \frac{\log(\tau)}{\tau^{-\zeta}} \overset{\text{L'H}}{=} \lim_{\tau \to 0} \frac{\frac{1}{\tau}}{-\zeta \cdot \tau^{-\zeta-1}} = \lim_{\tau \to 0} \frac{\tau^\zeta}{-\zeta} = 0$$
∎

We will need the following operators:

Definition 1 (Interior Push). Given $x \in \mathcal{X}$, define $\bar{x}$ and $\check{x}$ as a modified x whose components z have been pushed by an amount into the interior if they are close to zero:

$$\bar{z}_{[j]}(t) := \max\left\{z_{[j]}(t), \tau/L_\omega\right\},$$
$$\check{z}_{[j]}(t) := \max\left\{z_{[j]}(t), \tau/(2 \cdot L_\omega)\right\}$$

for all $j \in \{1, 2, \ldots, n_z\}$ and $t \in \overline{\Omega}$.

Note that $\bar{x} \in \mathcal{X}$ and that $x_{\omega,\tau}^\star = \bar{x}_{\omega,\tau}^\star$ from Lemma 2.

We show that the barrier function is bounded for the elements that have been pushed into the interior.

Lemma 4 (Bound for Γ). *If $x \in \mathcal{X}$ with $\|z\|_{L^\infty(\Omega)} = \mathcal{O}(1)$, then*

$$|\tau \cdot \Gamma(\bar{x})| = \mathcal{O}\left(\tau^{1-\zeta}\right), \qquad |\tau \cdot \Gamma(\check{x})| = \mathcal{O}\left(\tau^{1-\zeta}\right).$$

Proof: Since the definitions are similar, we only show the proof for $\bar{x}$.

$$|\tau \cdot \Gamma(\bar{x})| \leq \left| \tau \cdot \sum_{j=1}^{n_z} \int_\Omega \log\big(\bar{z}_{[j]}(t)\big) \, \mathrm{d}t \right|$$
$$\leq n_z \cdot |\Omega| \cdot \max_{1 \leq j \leq n_z} \|\tau \cdot \log(\bar{z}_{[j]})\|_{L^\infty(\Omega)}$$
$$\leq n_z \cdot |\Omega| \cdot \Big(\underbrace{\mathcal{O}\left(\tau^{1-\zeta}\right)}_{\text{bound for } \bar{z}_{[j]} < 1} + \underbrace{\mathcal{O}(\tau)}_{\text{bound for } \bar{z}_{[j]} \geq 1} \Big)$$
$$= \mathcal{O}\left(\tau^{1-\zeta}\right).$$

In the third line, we have distinguished two cases, namely $\left|\log\big(\bar{z}_{[j]}(t)\big)\right|$ attains its essential supremum at a $t \in \overline{\Omega}$ where either $\bar{z}_{[j]}(t) < 1$ (case 1) or where $\bar{z}_{[j]}(t) \geq 1$ (case 2). In the first case, we can use Lemma 3. In the second case, we simply bound the logarithm using $\|\bar{z}_{[j]}\|_{L^\infty(\Omega)} \leq \|z\|_{L^\infty(\Omega)} = \mathcal{O}(1)$ to arrive at the term $\mathcal{O}(\tau)$. ∎

We can use this result to show below that $x_{\omega,\tau}^\star$ is ε-optimal for (PP) if the L^∞-norms of $z_\omega^\star$ and $z_{\omega,\tau}^\star$ are bounded.

Proposition 2 (Penalty-Barrier Solution). *If* $\|z_\omega^\star\|_{L^\infty(\Omega)}, \|z_{\omega,\tau}^\star\|_{L^\infty(\Omega)} = \mathcal{O}(1)$, *then*

$$|F_\omega(x_{\omega,\tau}^\star) - F_\omega(x_\omega^\star)| = \mathcal{O}\left(\tau^{1-\zeta}\right).$$

Proof: From the definition of the bar-operator, we can use the bound

$$\|x_\omega^\star - \bar{x}_\omega^\star\|_{\mathcal{X}} = \|z_\omega^\star - \bar{z}_\omega^\star\|_{L^2(\Omega)} = \sqrt{\int_\Omega \|z_\omega^\star - \bar{z}_\omega^\star\|_2^2 \, dt}$$

$$\leq \sqrt{|\Omega| \cdot n_z \cdot \|z_\omega^\star - \bar{z}_\omega^\star\|_{L^\infty(\Omega)}^2} \leq n_z \cdot \sqrt{|\Omega|} \cdot \frac{\tau}{L_\omega},$$

together with the facts that $x_{\omega,\tau}^\star = \bar{x}_{\omega,\tau}^\star$ and F_ω is Lipschitz continuous, to get

$$0 \leq F_\omega(x_{\omega,\tau}^\star) - F_\omega(x_\omega^\star)$$
$$\leq F_\omega(\bar{x}_{\omega,\tau}^\star) - F_\omega(\bar{x}_\omega^\star) + L_\omega \cdot \|x_\omega^\star - \bar{x}_\omega^\star\|_{\mathcal{X}}$$
$$\leq \underbrace{F_\omega(\bar{x}_{\omega,\tau}^\star) - F_{\omega,\tau}(\bar{x}_{\omega,\tau}^\star)}_{= -\tau \cdot \Gamma(\bar{x}_{\omega,\tau}^\star)} + F_{\omega,\tau}(\bar{x}_{\omega,\tau}^\star)$$
$$- \left(\underbrace{F_\omega(\bar{x}_\omega^\star) - F_{\omega,\tau}(\bar{x}_\omega^\star)}_{= -\tau \cdot \Gamma(\bar{x}_\omega^\star)} + F_{\omega,\tau}(\bar{x}_\omega^\star) \right) + L_\omega \cdot n_z \cdot \sqrt{|\Omega|} \cdot \frac{\tau}{L_\omega}$$
$$\leq F_{\omega,\tau}(\bar{x}_{\omega,\tau}^\star) - F_{\omega,\tau}(\bar{x}_\omega^\star)$$
$$+ |\tau \cdot \Gamma(\bar{x}_{\omega,\tau}^\star)| + |\tau \cdot \Gamma(\bar{x}_\omega^\star)| + n_z \cdot \sqrt{|\Omega|} \cdot \tau.$$

For the terms $|\tau \cdot \Gamma(\bar{x}_{\omega,\tau}^\star)|$ and $|\tau \cdot \Gamma(\bar{x}_\omega^\star)|$ we use Lemma 4 to obtain the result from

$$F_\omega(x_{\omega,\tau}^\star) - F_\omega(x_\omega^\star) \leq \underbrace{F_{\omega,\tau}(\bar{x}_{\omega,\tau}^\star) - F_{\omega,\tau}(\bar{x}_\omega^\star)}_{\leq 0}$$
$$+ \mathcal{O}\left(\tau^{1-\zeta}\right) + n_z \cdot \sqrt{|\Omega|} \cdot \tau.$$

The under-braced term is bounded above by zero because $\bar{x}_{\omega,\tau}^\star = x_{\omega,\tau}^\star$ is a minimizer of $F_{\omega,\tau}$. ∎

Remark 1. Prop. 2 uses (A.4), i.e. $\|z_\omega^\star\|_{L^\infty(\Omega)}, \|z_{\omega,\tau}^\star\|_{L^\infty(\Omega)} = \mathcal{O}(1)$. Note that the assumption can be enforced. For example, the path constraints

$$z_{[1]}(t) \geq 0, \quad z_{[2]}(t) \geq 0, \quad z_{[1]}(t) + z_{[2]}(t) = const$$

lead to $\|z_{[j]}\|_{L^\infty(\Omega)} \leq const, \forall j \in \{1,2\}$. Constraints like these arise when variables have simple upper and lower bounds before being transformed into the form (DOP).

Similarly, boundedness of $\|x\|_{\mathcal{X}}$ can be enforced. To this end, introduce box constraints for each component of $\dot{y}, y, z$. This can be done by using inequality path constraints c_i, described in Appendix A, and then transforming these constraints into the form (DOP).

III. Finite Element Method

Our method constructs x_h^ϵ by solving the unconstrained problem (PBP) computationally in a finite-dimensional subspace of $\mathcal{X}$, using nonlinear optimization methods. The subspace is constructed using the Finite Element Method.
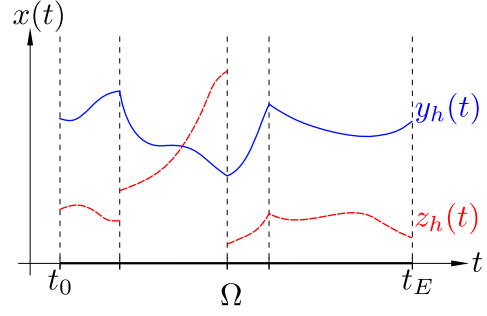

Fig. 1. Continuous and discontinuous piecewise polynomial finite element functions y_h, z_h on a mesh $\mathcal{T}_h$ of four intervals.

In this section we introduce a suitable finite-dimensional space. We then show a stability result. Eventually, we prove convergence of the Finite Element solution to solutions of (PBP) and (DOP).

A. Definition of the Finite Element Space

Let the mesh parameter $h \in (0, |\Omega|]$. The set $\mathcal{T}_h$ is called a *mesh* and consists of open intervals $T \subset \Omega$ that satisfy the usual conditions [29, Chap. 2]:

(i) Disjunction: $T_1 \cap T_2 = \emptyset$ $\forall$ distinct $T_1, T_2 \in \mathcal{T}_h$.
(ii) Coverage: $\bigcup_{T \in \mathcal{T}_h} \overline{T} = \overline{\Omega}$.
(iii) Resolution: $\max_{T \in \mathcal{T}_h} |T| = h$.
(iv) Quasi-uniformity: $\min_{T_1, T_2 \in \mathcal{T}_h} \frac{|T_1|}{|T_2|} \geq \sigma > 0$.

The constant σ must not depend on h and $1/\sigma = \mathcal{O}(1)$.

We write $\mathcal{P}_p(T)$ for the space of functions that are polynomials of degree $\leq p \in \mathbb{N}_0$ on interval T. Our Finite Element space is then given as

$$\mathcal{X}_{h,p} := \left\{ x : \overline{\Omega} \to \mathbb{R}^{n_x} \mid y \in \mathcal{C}^0(\overline{\Omega}), x \in \mathcal{P}_p(T) \, \forall T \in \mathcal{T}_h \right\}.$$

$\mathcal{X}_{h,p} \subset \mathcal{X}$ is a Hilbert-space with scalar product $\langle \cdot, \cdot \rangle_{\mathcal{X}}$.

Note that if $(y, z) \in \mathcal{X}_{h,p}$, then y is continuous but $\dot{y}$ and z can be discontinuous. Figure 1 illustrates two functions y_h, z_h from the finite elements spaces. Both functions are piecewise polynomials over each interval $T \in \mathcal{T}_h$. However, y_h is continuous, whereas z_h can have jumps between the intervals.

B. Discrete Penalty-Barrier Problem

We state the *discrete penalty-barrier problem* as

$$\text{Find } x_h^\star \in \underset{x \in \mathcal{X}_{h,p}^{\omega,\tau}}{\arg\min} \, F_{\omega,\tau}(x) \qquad \text{(PBP}_h\text{)}$$

where the space

$$\mathcal{X}_{h,p}^{\omega,\tau} := \left\{ x \in \mathcal{X}_{h,p} \mid z(t) \geq \tau/(2 \cdot L_\omega) \cdot \mathbf{1} \text{ f.a.e. } t \in \Omega \right\}.$$

Note that Lemma 2 applies to solutions to (PBP), whereas we will consider sub-optimal solutions to (PBP$_h$) below, hence we cannot guarantee that these sub-optimal solutions will satisfy $z(t) \geq \tau/(1 \cdot L_\omega) \cdot \mathbf{1}$. The looser constraint $z(t) \geq \tau/(2 \cdot L_\omega) \cdot \mathbf{1}$ in the definition above will be used in the proof of Theorem 2 below. See also Section V for a brief discussion on why the latter constraints can be omitted in a practical numerical method.

C. Stability

The following result shows that two particular Lebesgue norms are equivalent in the above Finite Element space.

Lemma 5 (Norm equivalence). *If $x \in \mathcal{X}_{h,p}$, then*

$$\|x_{[j]}\|_{L^\infty(\Omega)} \leq \frac{p+1}{\sqrt{\sigma \cdot h}} \cdot \|x\|_{\mathcal{X}} \quad \forall j \in \{1, 2, \ldots, n_x\}.$$

Proof: We can bound $\|x_{[j]}\|_{L^\infty(\Omega)} \leq \max_{T \in \mathcal{T}_h} \|x_{[j]}\|_{L^\infty(T)}$. We now use (10) in Appendix B. Since $x_{[j]} \in \mathcal{P}_p(T)$, it follows that

$$\max_{T \in \mathcal{T}_h} \|x_{[j]}\|_{L^\infty(T)} \leq \max_{T \in \mathcal{T}_h} \frac{p+1}{\sqrt{|T|}} \cdot \|x_{[j]}\|_{L^2(T)}$$
$$\leq \frac{p+1}{\sqrt{\sigma \cdot h}} \cdot \|x_{[j]}\|_{L^2(\Omega)} \leq \frac{p+1}{\sqrt{\sigma \cdot h}} \cdot \|x\|_{\mathcal{X}}.$$

∎

We can now obtain a bound on the growth of $F_{\omega,\tau}$ in a neighborhood of a solution $x^\star_{\omega,\tau}$ to (PBP) for elements in $\mathcal{X}_{h,p}$.

Proposition 3 (Lipschitz continuity). *Let*

$$\delta_{\omega,\tau,h} := \frac{\tau}{2 \cdot L_\omega} \cdot \frac{\sqrt{\sigma \cdot h}}{p+1},$$

$$L_{\omega,\tau,h} := L_\omega + n_z \cdot |\Omega| \cdot 2 \cdot L_\omega \cdot \frac{p+1}{\sqrt{\sigma \cdot h}}.$$

Consider the spherical neighbourhood

$$\mathcal{B} := \left\{ x \in \mathcal{X} \mid \|x^\star_{\omega,\tau} - x\|_{\mathcal{X}} \leq \delta_{\omega,\tau,h} \right\}.$$

The following holds $\forall x^A, x^B \in \mathcal{B} \cap \mathcal{X}_{h,p}$:

$$|F_{\omega,\tau}(x^A) - F_{\omega,\tau}(x^B)| \leq L_{\omega,\tau,h} \cdot \|x^A - x^B\|_{\mathcal{X}}.$$

Proof: From Lemma 2 and Lemma 5 it follows that

$$\operatorname*{ess\,inf}_{t \in \Omega} z_{[j]}(t) \geq \underbrace{\operatorname*{ess\,inf}_{t \in \Omega} z^\star_{\omega,\tau,[j]}(t)}_{\geq \frac{\tau}{L_\omega}} - \underbrace{\|z^\star_{\omega,\tau,[j]} - z_{[j]}\|_{L^\infty(\Omega)}}_{\leq \frac{p+1}{\sqrt{\sigma \cdot h}} \cdot \delta_{\omega,\tau,h} \leq \frac{\tau}{2 \cdot L_\omega}} \cdot$$

holds $\forall x \in \mathcal{B} \cap \mathcal{X}_{h,p}$. Hence,

$$\min_{1 \leq j \leq n_z} \operatorname*{ess\,inf}_{t \in \Omega} z_{[j]}(t) \geq \frac{\tau}{2 \cdot L_\omega} \quad \forall x \in \mathcal{B} \cap \mathcal{X}_{h,p}. \quad (3)$$

From Lipschitz-continuity of F_ω we find

$$|F_{\omega,\tau}(x^A) - F_{\omega,\tau}(x^B)| \leq |F_\omega(x^A) - F_\omega(x^B)|$$
$$+ \tau \cdot \sum_{j=1}^{n_z} \int_\Omega \left| \log\left(z^A_{[j]}(t)\right) - \log\left(z^B_{[j]}(t)\right) \right| \, \mathrm{d}t$$
$$\leq L_\omega \cdot \|x^A - x^B\|_{\mathcal{X}} + \tau \cdot n_z \cdot |\Omega|$$
$$\cdot \max_{1 \leq j \leq n_z} \operatorname*{ess\,sup}_{t \in \Omega} \left| \log\left(z^A_{[j]}(t)\right) - \log\left(z^B_{[j]}(t)\right) \right|.$$

We know a lower bound for the arguments of the logarithm from (3). Thus, the essential supremum term can be bounded with a Lipschitz result for the logarithm:

$$\max_{1 \leq j \leq n_z} \operatorname*{ess\,sup}_{t \in \Omega} \left| \log\left(z^A_{[j]}(t)\right) - \log\left(z^B_{[j]}(t)\right) \right|$$
$$\leq \max_{1 \leq j \leq n_z} \frac{1}{\frac{\tau}{2 \cdot L_\omega}} \cdot \|z^A_{[j]} - z^B_{[j]}\|_{L^\infty(\Omega)}$$
$$\leq \frac{2 \cdot L_\omega}{\tau} \cdot \frac{p+1}{\sqrt{\sigma \cdot h}} \cdot \|x^A - x^B\|_{\mathcal{X}},$$

where the latter inequality is obtained using Lemma 5. ∎

D. Interpolation Error

In order to show high-order convergence results, it is imperative that the solution function can be represented with high accuracy in a finite element space. In the following we introduce a suitable assumption for this purpose.

Motivated by the Bramble-Hilbert Lemma [30], we make the assumption (A.5) that for a fixed chosen degree $p = \mathcal{O}(1)$ there exists an $\ell \in (0, \infty)$ such that

$$\min_{x_h \in \mathcal{X}_{h,p}} \|x^\star_{\omega,\tau} - x_h\|_{\mathcal{X}} = \mathcal{O}\left(h^{\ell+1/2}\right). \quad (4)$$

Notice that the minimizer exists since $\mathcal{X}_{h,p}$ is a Hilbert space with induced norm $\|\cdot\|_{\mathcal{X}}$. In Appendix C we give two examples to demonstrate the mildness of assumption (4).

For the remainder, we define $\nu := \ell/2$, $\eta := (1 - \zeta) \cdot \nu$ with respect to ℓ, ζ. We choose $\tau = \mathcal{O}(h^\nu)$ and $\omega = \mathcal{O}(h^\eta)$ with $h > 0$ suitably small such that $0 < \tau \leq \omega \leq 0.5$.

Following the assumption (4), the result below shows that the best approximation in the finite element space satisfies an approximation property.

Lemma 6 (Finite Element Approximation Property). *If* (4) *holds and $h > 0$ is chosen sufficiently small, then*

$$\min_{x_h \in \mathcal{X}_{h,p}} \|x^\star_{\omega,\tau} - x_h\|_{\mathcal{X}} \leq \delta_{\omega,\tau,h}. \quad (5)$$

Proof: For $h > 0$ sufficiently small it follows from $\ell > \nu + \eta$, that $h^{\ell+1/2} < h^{\nu+\eta+1/2}$. Hence,

$$\min_{x_h \in \mathcal{X}_{h,p}} \|x^\star_{\omega,\tau} - x_h\|_{\mathcal{X}} \leq const \cdot h^{\nu+\eta+1/2}$$

for some constant. Note that

$$\delta_{\omega,\tau,h} \geq \frac{\tau}{\frac{L_r}{\omega}} \cdot \frac{\sqrt{\sigma \cdot h}}{p+1} = \frac{\sqrt{\sigma}}{L_r \cdot (p+1)} \cdot \tau \cdot \omega \cdot \sqrt{h}$$
$$\geq const \cdot h^{\nu+\eta+1/2}.$$

The result follows. ∎

In other words, Lemma 6 says for $h > 0$ sufficiently small it follows that $\mathcal{B} \cap \mathcal{X}_{h,p} \neq \emptyset$. This is because the minimizing argument of (5) is an element of $\mathcal{B}$.

E. Optimality

We show that an ϵ-optimal solution for (PBP$_h$) is an ε-optimal solution for (PBP), where $\varepsilon \geq \epsilon$.

Theorem 1 (Optimality of Unconstrained FEM Minimizer). *Let $\mathcal{B}$ be as in Proposition 3. Let x_h^ϵ be an ϵ-optimal solution for* (PBP$_h$)*, i.e.*

$$F_{\omega,\tau}(x_h^\epsilon) \leq F_{\omega,\tau}(x_h^\star) + \epsilon.$$

If $\mathcal{B} \cap \mathcal{X}_{h,p} \neq \emptyset$, then x_h^ϵ satisfies

$$F_{\omega,\tau}(x_h^\epsilon) \leq F_{\omega,\tau}(x^\star_{\omega,\tau}) + \epsilon + L_{\omega,\tau,h} \cdot \min_{x_h \in \mathcal{X}_{h,p}} \|x^\star_{\omega,\tau} - x_h\|_{\mathcal{X}}.$$

Proof: Consider the unique Finite Element best approximation from (5)

$$\tilde{x}_h := \operatorname*{arg\,min}_{x_h \in \mathcal{X}_{h,p}} \|x^\star_{\omega,\tau} - x_h\|_{\mathcal{X}},$$

Since $\mathcal{B} \cap \mathcal{X}_{h,p} \neq \emptyset$ by assumption, it follows $\tilde{x}_h \in \mathcal{B} \cap \mathcal{X}_{h,p}$. Hence,

$$\tilde{x}_h = \operatorname*{arg\,min}_{x_h \in \mathcal{B} \cap \mathcal{X}_{h,p}} \|x^\star_{\omega,\tau} - x_h\|_{\mathcal{X}}.$$

From (3) we find $\mathcal{B} \cap \mathcal{X}_{h,p} \subset \mathcal{X}^{\omega,\tau}_{h,p}$. Thus, $\tilde{x}_h \in \mathcal{X}^{\omega,\tau}_{h,p} \subset \mathcal{X}_{h,p}$. Hence,

$$\tilde{x}_h = \operatorname*{arg\,min}_{x_h \in \mathcal{X}^{\omega,\tau}_{h,p}} \|x^\star_{\omega,\tau} - x_h\|_{\mathcal{X}}.$$

Proposition 3 can be used to obtain the bound

$$F_{\omega,\tau}(\tilde{x}_h) \leq F_{\omega,\tau}(x^\star_{\omega,\tau}) + L_{\omega,\tau,h} \cdot \|x^\star_{\omega,\tau} - \tilde{x}_h\|_{\mathcal{X}}.$$

Since $x^\star_{\omega,\tau}$ is a global minimizer of $F_{\omega,\tau}$ in $\mathcal{X}$, x^ϵ_h is a global ϵ-optimal minimizer of $F_{\omega,\tau}$ in the subspace $\mathcal{X}^{\omega,\tau}_{h,p} \subset \mathcal{X}$ and $\tilde{x}_h$ lives in $\mathcal{X}^{\omega,\tau}_{h,p}$, one can show that

$$F_{\omega,\tau}(x^\star_{\omega,\tau}) \leq F_{\omega,\tau}(x^\epsilon_h) \leq F_{\omega,\tau}(\tilde{x}_h) + \epsilon.$$

The result follows from the above. ∎

F. Convergence

We obtain a bound for the optimality gap and feasibility residual of x^ϵ_h.

Theorem 2 (Convergence to (DOP)). *Let x^ϵ_h be an ϵ-optimal numerical solution to* (PBP$_h$). *If* $\|z^\star_\omega\|_{L^\infty(\Omega)}$, $\|z^\star_{\omega,\tau}\|_{L^\infty(\Omega)}$, $\|z^\epsilon_h\|_{L^\infty(\Omega)} = \mathcal{O}(1)$, *then x^ϵ_h satisfies*

$$g_{\mathrm{opt}} = \mathcal{O}\left(\tau^{1-\zeta} + \varepsilon_{h,p}\right), \ r_{\mathrm{feas}} = \mathcal{O}\left(\omega \cdot \left(1 + \tau^{1-\zeta} + \varepsilon_{h,p}\right)\right),$$

where

$$\varepsilon_{h,p} := L_{\omega,\tau,h} \cdot \min_{x_h \in \mathcal{X}_{h,p}} \|x^\star_{\omega,\tau} - x_h\|_{\mathcal{X}} + \epsilon.$$

Proof: From Theorem 1 we know $F_{\omega,\tau}(x^\epsilon_h) \leq F_{\omega,\tau}(x^\star_{\omega,\tau}) + \varepsilon_{h,p}$. This is equivalent to

$$F_\omega(x^\epsilon_h) + \tau \cdot \Gamma(x^\epsilon_h) \leq F_\omega(x^\star_{\omega,\tau}) + \tau \cdot \Gamma(x^\star_{\omega,\tau}) + \varepsilon_{h,p}$$
$$\Rightarrow F_\omega(x^\epsilon_h) \leq F_\omega(x^\star_{\omega,\tau}) + \underbrace{|\tau \cdot \Gamma(x^\epsilon_h)| + |\tau \cdot \Gamma(x^\star_{\omega,\tau})|}_{(*)} + \varepsilon_{h,p}.$$

Since $x^\epsilon_h \in \mathcal{X}^{\omega,\tau}_{h,p}$, it follows that $z^\epsilon_h \geq \frac{\tau}{2 \cdot L_\omega} \cdot \mathbf{1}$ and thus $x^\epsilon_h = \check{x}^\epsilon_h$. From Lemma 2 we know $x^\star_{\omega,\tau} = \check{\tilde{x}}^\star_{\omega,\tau}$. Thus, we can apply Lemma 4 to bound $(*)$ with $\mathcal{O}(\tau^{1-\zeta})$. It follows that

$$F_\omega(x^\epsilon_h) \leq F_\omega(x^\star_{\omega,\tau}) + \mathcal{O}(\tau^{1-\zeta}) + \varepsilon_{h,p}.$$

Since, according to Proposition 2, $x^\star_{\omega,\tau}$ is $\tilde{\varepsilon}$-optimal for (PP), where $\tilde{\varepsilon} = \mathcal{O}(\tau^{1-\zeta})$, it follows that

$$F_\omega(x^\epsilon_h) \leq F_\omega(x^\star_\omega) + \underbrace{\mathcal{O}(\tau^{1-\zeta}) + \varepsilon_{h,p}}_{=: \varepsilon}.$$

In other words, x^ϵ_h is ε-optimal for (PP). The result now follows from Proposition 1. ∎

Below, we translate the above theorem into an order of convergence result. For piecewise smooth optimal control solutions, ℓ is usually in the order of p.

Theorem 3 (Order of Convergence to (DOP)). *Consider x^ϵ_h with $\epsilon = \mathcal{O}(h^{\ell-\eta})$, and let* (4) *hold. Then:*

$$g_{\mathrm{opt}} = \mathcal{O}\left(h^\eta\right), \qquad r_{\mathrm{feas}} = \mathcal{O}\left(h^\eta\right).$$

Proof: It holds that

$$\begin{aligned} L_{\omega,\tau,h} =& \ \ L_\omega \ \ + (n_z \cdot |\Omega| \cdot 2) \cdot \ L_\omega \ \ \cdot \left(\frac{p+1}{\sqrt{\sigma \cdot h}}\right) \\ =& \frac{\mathcal{O}(1)}{\omega} + \ \ \ \mathcal{O}(1) \ \ \cdot \frac{\mathcal{O}(1)}{\omega} \cdot \frac{\mathcal{O}(1)}{\sqrt{h}} \\ =& \ \mathcal{O}\left(h^{-\eta} + h^{-\eta} \cdot h^{-1/2}\right) = \mathcal{O}\left(h^{-\eta-1/2}\right). \end{aligned}$$

From (4), we find

$$\begin{aligned} \varepsilon_{h,p} =& L_{\omega,\tau,h} \cdot \min_{x_h \in \mathcal{X}_{h,p}} \|x^\star_{\omega,\tau} - x_h\|_{\mathcal{X}} + \epsilon \\ =& \mathcal{O}\left(h^{-\eta-1/2}\right) \cdot \mathcal{O}\left(h^{\ell+1/2}\right) + \mathcal{O}(h^{\ell-\eta}) = \mathcal{O}\left(h^{\ell-\eta}\right). \end{aligned}$$

Combining the result for $\varepsilon_{h,p}$ with Theorem 2, we find x^ϵ_h satisfies

$$\begin{aligned} g_{\mathrm{opt}} =& \mathcal{O}\left(\tau^{1-\zeta} + \varepsilon_{h,p}\right) \\ =& \mathcal{O}\left(h^{\nu \cdot (1-\zeta)} + h^{\ell-\eta}\right) = \mathcal{O}\left(h^{\min\{\eta,\,\ell-\eta\}}\right), \\ r_{\mathrm{feas}} =& \mathcal{O}\left(\omega \cdot (1 + \tau^{(1-\zeta) \cdot \nu} + \varepsilon_{h,p})\right) \\ =& \mathcal{O}\left(h^\eta + h^{\eta + (1-\zeta) \cdot \nu} + h^{\eta + \ell - \eta}\right) \\ =& \mathcal{O}\left(h^\eta + h^{2 \cdot \eta} + h^\ell\right) = \mathcal{O}\left(h^{\min\{\eta,\,\ell\}}\right). \end{aligned}$$

Note that $\ell > \ell - \eta > \eta$. ∎

Recall that $\eta = \ell/2 \cdot (1 - \zeta)$, where $0 < \zeta \ll 1$. If $\ell \approx p$ and $\eta \approx \ell/2$ it follows that $h^\eta \approx \sqrt{h^p}$.

G. Numerical Quadrature

When computing x^ϵ_h, usually the integrals in F and r cannot be evaluated exactly. In this case, one uses numerical quadrature and replaces $F_{\omega,\tau}$ with

$$F_{\omega,\tau,h} := F_h + \frac{1}{2 \cdot \omega} \cdot r_h + \tau \cdot \Gamma.$$

Since $\mathcal{X}_{h,p}$ is a space of piecewise polynomials, Γ can be integrated analytically. However, the analytic integral expressions become very complicated. This is why, for a practical method, one may also wish to use quadrature for Γ.

If F and r have been replaced with quadrature approximations F_h, r_h, then it is sufficient that these approximations satisfy

$$|F_{\omega,\tau,h}(x) - F_{\omega,\tau}(x)| \leq C_{\mathrm{quad}} \cdot \frac{h^q}{\omega} \qquad \forall x \in \mathcal{X}_{h,p}, \quad (6)$$

with a constant $C_{\mathrm{quad}} \in (0, \infty)$ and $q \in \mathbb{N}$ the quadrature order, to ensure that the convergence theory holds, as we show below.

The quadrature error can be bounded independent of x since $|f|$ and $\|c\|_1$ are bounded globally. (6) poses a consistency condition and a stability condition.

a) Consistency: There is a consistency condition in (6) that relates to suitable values of q. In particular, if we want to make sure to converge of order $\mathcal{O}(h^\eta)$, as presented in Theorem 3, then q has to be sufficiently large. Consider the problem

$$\tilde{x}^\star_h \in \operatorname*{arg\,min}_{x \in \mathcal{X}^{\omega,\tau}_{h,p}} \ F_{\omega,\tau,h}(x).$$

Note that $\tilde{x}_h^\star$ is ϵ-optimal for (PBP_h), where from

$$F_{\omega,\tau}(\tilde{x}_h^\star) - C_{\mathrm{quad}} \cdot \frac{h^q}{\omega} \leq F_{\omega,\tau,h}(\tilde{x}_h^\star) \leq F_{\omega,\tau,h}(x_h^\star)$$
$$\leq F_{\omega,\tau}(x_h^\star) + C_{\mathrm{quad}} \cdot \frac{h^q}{\omega}$$

it follows that $\epsilon = \mathcal{O}(h^q/\omega) = \mathcal{O}(h^{q-\eta})$. Hence, $\tilde{x}_h^\star$ satisfies the bounds for the optimality gap and feasibility residual presented in Theorem 2. We obtain the same order of convergence as in Theorem 3 when maintaining $\epsilon = \mathcal{O}(h^{\ell-\eta})$, i.e. choosing $q \geq \ell$.

b) Stability: Beyond consistency, (6) poses a non-trivial stability condition. This is because the error bound must hold $\forall x \in \mathcal{X}_{h,p}$. We illustrate this with an example.

Consider $\Omega = (0,1)$, $n_y = 0$, $n_z = 1$, $c(x) := \sin(\pi \cdot x)$, i.e. the constraint forces $x(t) = 0$. Clearly, $\|c\|_1$ and $\|\nabla c\|_1$ are bounded globally. We use the uniform mesh $\mathcal{T}_h := \{ T_j \mid T_j = ((j-1)\cdot h, j\cdot h), j = 1, 2, \ldots, 1/h \}$ for $h \in 1/\mathbb{N}$. Use $\mathcal{X}_{h,1}$ for the Finite Element space and Gauss-Legendre quadrature of order 3, i.e. the mid-point rule quadrature scheme [31]. If the candidate $x \in \mathcal{X}_{h,1}$ is defined as $x(t) := -1/h + 2/h \cdot (t - j \cdot h)$ for $t \in T_j$, then the quadrature error is

$$|r_h(x) - r(x)|$$
$$= \left| \underbrace{h \cdot \sum_{j=1}^{1/h} \sin^2\big(\pi \cdot x(j \cdot h - h/2)\big)}_{=0} - \underbrace{\int_0^1 \sin^2\big(\pi \cdot x(t)\big)\mathrm{d}t}_{=0.5} \right|,$$

which violates (6). In contrast, using Gauss-Legendre quadrature of order 5 yields satisfaction of (6) with $q = 5$.

In order to satisfy (6), a suitable quadrature rule must take into account the polynomial degree p of the finite element space and the nature of the nonlinearity of c. If

$$\left(f + \frac{1}{2 \cdot \omega} \cdot \|c\|_2^2 \right) \circ x \in \mathcal{P}_d(T), \ \forall T \in \mathcal{T}_h, \forall x \in \mathcal{X}_{h,p} \cap \mathcal{B},$$

for some $d \in \mathbb{N}$ then it means that the integrals of F and r are polynomials in t. In this case, $q \geq d$ is a sufficient order of quadrature. For a practical method, we propose to use Gaussian quadrature of order $q = 4 \cdot p + 1$, i.e. using $2 \cdot p$ abscissae per interval $T \in \mathcal{T}_h$. We notice that this would not be possible with collocation, where the number of quadrature points cannot exceed the polynomial degree p of $\mathcal{X}_{h,p}$, as this would cause overdetermination of the nonlinear program.

IV. Solving the Nonlinear Program

We identify $x_h \in \mathcal{X}_{h,p}$ with a finite-dimensional vector $\mathbf{x} \in \mathbb{R}^{N_X}$ and describe the nonlinear optimization problem from which $\mathbf{x}$ can be computed numerically.

Using a quadrature rule of abscissae $s_j \in \overline{\Omega}$ and quadrature weights $\alpha_j \in \mathbb{R}_+$ for $j = 1, \ldots, n_q$, we can define the F_h and C_h as

$$F_h(\mathbf{x}) := \sum_{j=1}^{n_q} \alpha_j \cdot f\big(\dot{y}_h(s_j), y_h(s_j), z_h(s_j)\big),$$

$$C_h(\mathbf{x}) := \begin{pmatrix} b\big(y_h(t_1), y_h(t_2), \ldots, y_h(t_M)\big) \\ \sqrt{\alpha_1} \cdot c\big(\dot{y}_h(s_1), y_h(s_1), z_h(s_1)\big) \\ \sqrt{\alpha_2} \cdot c\big(\dot{y}_h(s_2), y_h(s_2), z_h(s_2)\big) \\ \vdots \\ \sqrt{\alpha_{n_q}} \cdot c\big(\dot{y}_h(s_{n_q}), y_h(s_{n_q}), z_h(s_{n_q})\big) \end{pmatrix}.$$

Note that $r_h(x_h) = \|C_h(\mathbf{x})\|_2^2$.

The optimization problem for $\mathbf{x}$ is

$$\min_{\mathbf{x} \in N_X} \quad \phi(\mathbf{x}) := F_h(\mathbf{x}) + \frac{1}{2 \cdot \omega} \cdot \|C_h(\mathbf{x})\|_2^2 + \tau \cdot \Gamma(\mathbf{x}). \quad (7)$$

In the formulation, care has to be taken to ensure that y is continuous, e.g. by appropriate choice of polynomial basis and coefficient or adding equality constraints between mesh intervals and including them in b. To avoid the need for distinction in a practical method, both y, z can be discretized using continuous finite elements only, as we have done in Section V. The additional inequality constraints $z(t) \geq \tau/(2 \cdot L_\omega) \cdot \mathbf{1}$ in (PBP_h) can be incorporated into the barrier function Γ as in (PBP). However, these constraints are usually inactive if h, τ, ω are sufficiently small and the computed point is close to $x_{\omega,\tau}^\star$, hence can be omitted in practice.

We briefly sketch how to minimize ϕ efficiently if it is sufficiently differentiable. The function ϕ has a structure similar to the merit functions used in sequential unconstrained minimization techniques [32]; there are quadratic penalty terms in ω and barrier terms in τ. If f, c, b are sufficiently differentiable, then the Jacobian and Hessian matrices of F_h, C_h are narrow-banded and of the same structure as for collocation methods [1, Chap. 4.6], with dimensions proportional to the number of mesh intervals times $n_c \cdot p$ and bandwidth proportional to $n_x \cdot p$. Efficient methods for computing these derivatives are described in [1, Chap 2.2], for example.

F_h is usually well-scaled. In contrast, the quadratic penalty term with C_h can cause issues, since for fine meshes and high orders ω is typically very small. A quadratic penalty method [33] or a modified version of the augmented Lagrangian method [34], [35] can be used to treat this term efficiently. The referenced methods are suitable to exploit sparsity of derivatives when solving large-scale problems.

Another term that requires careful treatment is the function Γ, which has roughly a similar numerical behaviour as a logarithmic barrier term. This can be handled efficiently using an interior-point strategy [36], i.e. the problem is initially solved for a larger value of τ, which is subsequently reduced in an iterative fashion.

A method combining both approaches is discussed in [37]. This is a primal-dual interior-point method, global-

ized with a line-search and a merit-function. Their merit-function has a similar structure to ϕ, using quadratic penalty terms and logarithmic barriers.

The choice of ω, τ deserves a brief discussion. As the mesh is refined, the numerical solution will become more accurate and eventually converge to some approximate solution. For a practical high-order method, we have used $\omega = \tau = 10^{-10}$ by default to obtain a sufficiently accurate solution. A more accurate solution can be obtained by decreasing ω, τ and refining the mesh even further. However, in practice, we usually do not use meshes so fine that we need to keep decreasing ω, τ in order to obtain a solution that satisfies our tolerance.

V. Numerical Results

We compare the convergence behaviour of our new penalty-barrier finite element method (PBF) against the state-of-the-art for direct transcription methods, that is, direct collocation on Legendre-Gauss-Radau (LGR) points [1], [3], [5], i.e. the Radau IIA Runge-Kutta method [38, p. 199]. As discussed earlier, collocation methods can encounter numerical difficulties when converging to control solutions on singular arcs. Convergence problems can also arise when constraints contain high-index DAEs. We demonstrate that our method does not suffer from any of these difficulties. To show this, we consider test problems from the literature that have bang-singular controls, totally singular controls and a high-index DAE.

Solutions presented for collocation methods were computed using ICLOCS2[1] [39] with the NLP solver IPOPT [27]. Solutions computed using the PBF method were obtained using the NLP solver described in [37], suitably adapted such that their primal merit function matches ϕ in (7).

The test problems have unique minimizers. The optimal control inputs are known in terms of either analytic expressions or highly accurate numerical representations.

A. Van der Pol Controller

This problem from [40] uses a controller to stabilize the van der Pol differential equations on a finite-time horizon. The problem is stated as

$$\min_{y,u} \quad \frac{1}{2} \cdot \int_0^4 \left(y_1(t)^2 + y_2(t)^2 \right) \, dt,$$
$$\text{s.t.} \quad y_1(0) = 0, \qquad y_2(0) = 1,$$
$$\dot{y}_1(t) = y_2(t),$$
$$\dot{y}_2(t) = -y_1(t) + y_2(t) \cdot \left(1 - y_1(t)^2 \right) + u(t),$$
$$-1 \le u(t) \le 1.$$

The problem features a bang-bang control with a singular arc on one sub-interval. The discontinuities in the optimal control are to five digits at $t_1 = 1.3667$ and $t_2 = 2.4601$.

We solved this problem with LGR collocation on 100 uniform elements of order 5. We compare this solution to

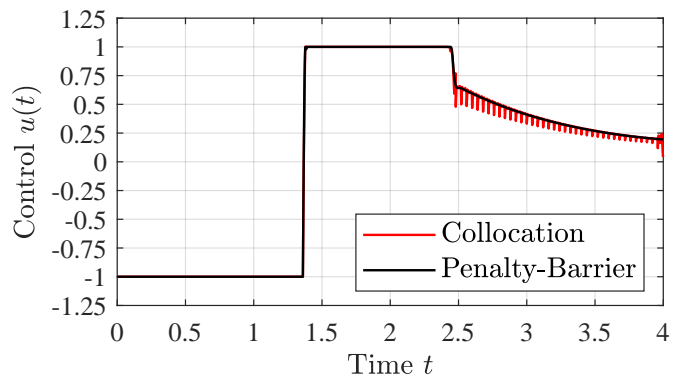[1]Downloadable from http://www.ee.ic.ac.uk/ICLOCS/

Fig. 2. Comparison of control input obtained with Penalty-Barrier method against LGR collocation for the Van der Pol problem.

the one obtained with PBF using 100 uniform elements of order $p = 5$, with $\omega = 10^{-10}$ and $\tau = 10^{-10}$.

Figure 2 presents the control profiles of the two numerical solutions. LGR shows ringing on the time interval $[t_2, 4]$ of the singular arc. In contrast, PBF converges to the analytic solution. The solution satisfies the error bounds $e(0) \approx 7.0 \cdot 10^{-2}$, $e(t_2) \approx 1.2 \cdot 10^{-2}$, $e(2.5) \approx 8.17 \cdot 10^{-4}$, and $e(2.6) \approx 9.6 \cdot 10^{-5}$, where $e(\hat{t}) := \|u^\star(t) - u_h(t)\|_{L^2([\hat{t},4])}$. The larger errors in the vicinity of the jumps occur due to the non-adaptive mesh.

B. Second-Order Singular Regulator

This bang-singular control problem from [18] is given as

$$\min_{y,u} \quad \frac{1}{2} \cdot \int_0^5 \left(y_1(t)^2 + y_2(t)^2 \right) \, dt,$$
$$\text{s.t.} \quad y_1(0) = 0, \quad y_2(0) = 1,$$
$$\dot{y}_1(t) = y_2(t), \quad \dot{y}_2(t) = u(t),$$
$$-1 \le u(t) \le 1.$$

Both LGR and PBF use 100 elements of order $p = 5$. PBF further uses $\omega = 10^{-10}$ and $\tau = 10^{-10}$. Figure 3 presents the control profiles of the two numerical solutions. LGR shows ringing on the time interval $[1.5, 5]$ of the singular arc. In contrast, PBF converges with the error $\|u^\star(t) - u_h(t)\|_{L^2([1.5, 5])} \approx 1.5 \cdot 10^{-4}$.

C. Aly-Chan Problem

The problem in [12], namely

$$\min_{y,u} \quad \frac{1}{2} \cdot \int_0^{\pi/2} \left(y_1(t)^2 - y_2(t)^2 \right) \, dt,$$
$$\text{s.t.} \quad y_1(0) = 0, \quad y_2(0) = 1,$$
$$\dot{y}_1(t) = y_2(t), \quad \dot{y}_2(t) = u(t),$$
$$-1 \le u(t) \le 1,$$

has a smooth totally singular control.

Both LGR and PBF use 100 elements of order $p = 5$. PBF further uses $\omega = 10^{-10}$ and $\tau = 10^{-10}$. Figure 4 presents the control profiles of the two numerical solutions. PBF converges, with error $\|u^\star(t) - u_h(t)\|_{L^2(\Omega)} \approx 3.7 \cdot 10^{-6}$. LGR does not converge for this problem, as was also observed in [11].
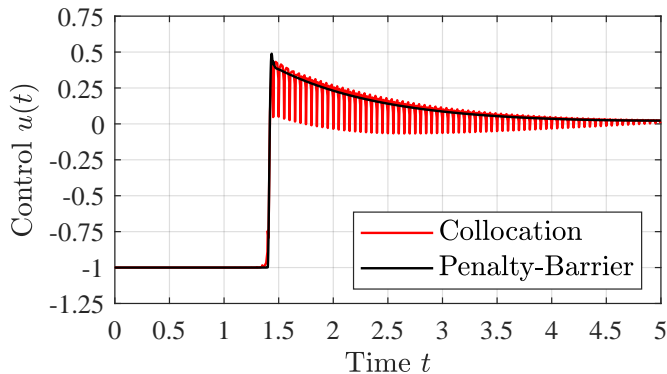
Fig. 3. Comparison of control input obtained with Penalty-Barrier method against LGR collocation for the second order singular regulator problem.
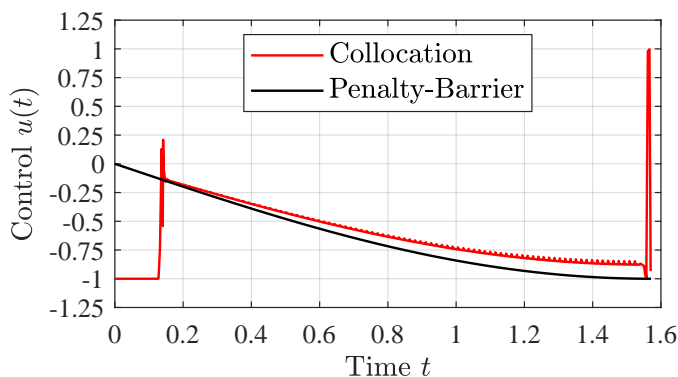

Fig. 4. Comparison of control input obtained with Penalty-Barrier method against LGR collocation for the Aly-Chan problem.

D. Pendulum in Differential-Algebraic Form

In this example from [41, Chap. 55], a control force decelerates a frictionless pendulum to rest. The objective is to minimize the integral of the square of the control:

$$\min_{\vec{\chi}, \xi, u} \qquad \int_0^3 u(t)^2 \, dt,$$
$$\text{s.t.} \quad \vec{\chi}(0) = (1, 0)^{\mathsf{T}}, \quad \dot{\vec{\chi}}(0) = \vec{0},$$
$$\vec{\chi}(3) = (0, -1)^{\mathsf{T}}, \quad \dot{\vec{\chi}}(3) = \vec{0},$$
$$\ddot{\vec{\chi}}(t) = (0, -9.81)^{\mathsf{T}} + 2 \cdot \vec{\chi}(t) \cdot \xi(t) + \vec{\chi}^{\perp}(t) \cdot u(t),$$

with an additional DAE constraint introduced below. In the constraints, the ODE for $\dot{\vec{\chi}}$ is a force balance in the pendulum mass. $\xi(t)$ is the beam force in the pendulum arm. $u(t)$ is the control force acting in the direction $\vec{\chi}^{\perp} := (-\chi_2, \chi_1)^{\mathsf{T}}$, $\vec{\chi}$ rotated by 90 degrees.

The DAE constraint is needed to determine the beam force in an implicit way, such that the pendulum arm has length 1 at any time; [41, Chap. 55] uses

$$0 = \|\dot{\vec{\chi}}(t)\|_2^2 - 2 \cdot \xi(t) - g \cdot \chi_2(t). \qquad (8)$$

The following alternative constraint achieves the same:

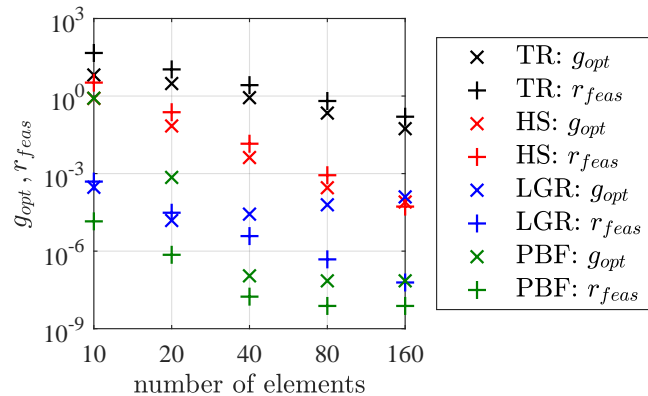$$0 = \|\vec{\chi}(t)\|_2^2 - 1. \qquad (8')$$

Fig. 5. Convergence of optimality gap and feasibility residual for Pendulum example, case A.

Notice that (8) is a DAE of index 1, whereas (8') is of index 3. Typically, a DAE of higher index is more difficult to solve numerically [38, Chap. VII.2].

In the following we study the convergence of four methods on meshes of increasing size: the Trapezoidal Method (TR) [5, Sec. 3], the Hermite-Simpson Method (HS) [5, Sec. 4], LGR and PBF. Both LGR and PBF use order $p = 5$. PBF further uses $\omega = 10^{-10}$ and $\tau = 10^{-10}$. We notice that TR/ HS have order $p = 3/ p = 3$, respectively. Our focus is primarily on determining whether a given method converges and only secondarily on orders of convergence. To find out where solvers struggle, we consider three experiments of the pendulum problem,

Case A where we consider the original problem with (8) as it is given in [41].

Case B where we add the path constraint $\xi(t) \leq 8$.

Case C where we exchange (8) in the original problem with (8').

All methods converge for case A. Figure 5 shows that TR converges slowly, while HS, LGR and PBF converge fast. At small magnitudes of $g_{\text{opt}}, r_{\text{feas}}$, further decrease of LGR and PBF deteriorates, presumably due to limits in solving the NLP accurately under rounding errors.

Case B is shown in Figure 6. The control force decelerates the pendulum more aggressively before the pendulum mass surpasses the lowest point, such that the beam force obeys the imposed upper bound. Figure 7 confirms convergence for all methods. The rate of convergence is slower compared to case A, as expected, because the solution of u is locally non-smooth. While the collocation methods tend to yield more accurate values for g_{opt} than r_{feas}, the feasibility residual of PBF is orders of magnitude smaller.

For case C, the collocation methods struggle: For HS on all meshes, the restoration phase in IPOPT converged to an infeasible point, indicating infeasibility of the discretized nonlinear program [27, Sec. 3.3]. For TR, the feasibility residual does not converge, as shown in Figure 8. Figure 9 shows that this is due to ringing in the numerical solution for the beam force. Regarding LGR, Figure 8 shows that the feasibility residual converges only for rela-
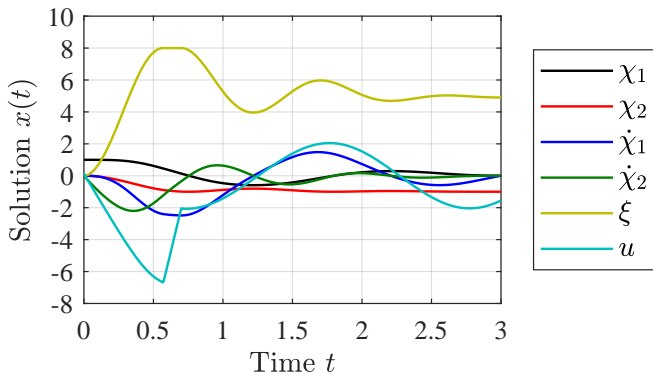
Fig. 6. Numerical solution of PBF on 80 elements for Pendulum example, case B.

Fig. 9. Numerical solutions of PBF and TR on 80 elements for Pendulum example, case C. The optimal control is identical to case A.
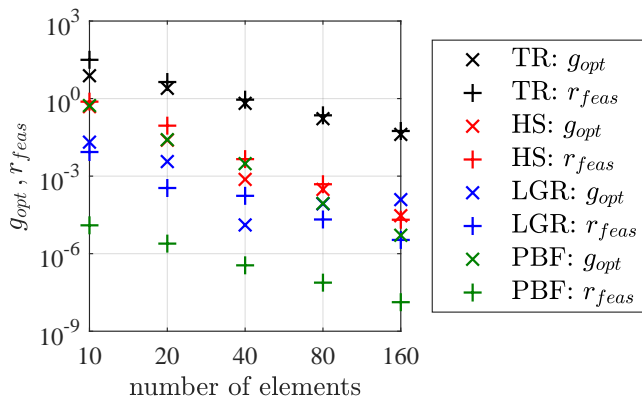

Fig. 7. Convergence of optimality gap and feasibility residual for Pendulum experiment, case B.

tively fine meshes. In contrast to the collocation methods, PBF does not show any signs of difficulty in convergence. That is, $g_{\text{opt}}, r_{\text{feas}}$ converge as fast as in case A.

VI. Conclusions

For the PBF method presented here we have proven convergence under mild and easily-enforced assumptions. Key to the convergence proof is the formulation of a suitable
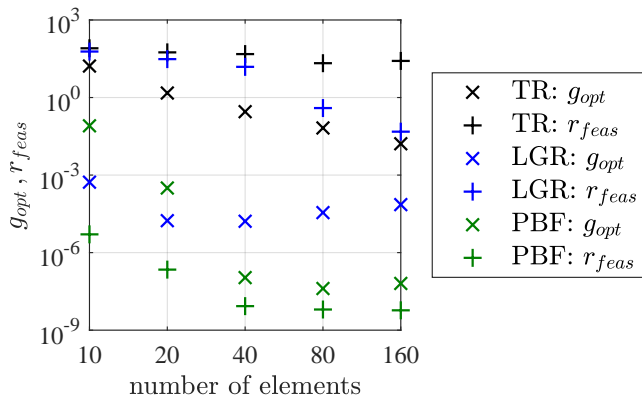

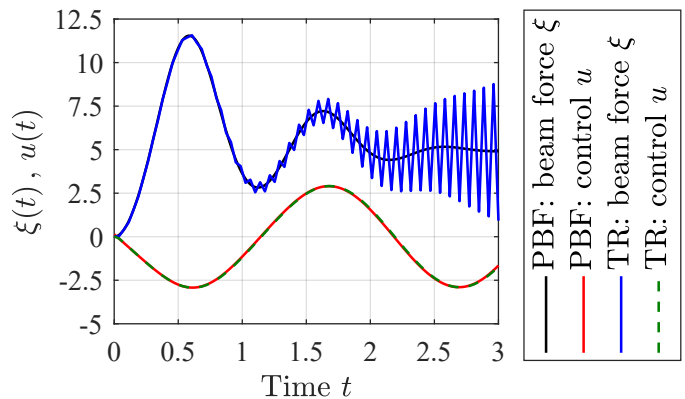Fig. 8. Convergence of optimality gap and feasibility residual for Pendulum example, case C.

unconstrained problem, which is discretized using finite elements. Since the $z^\star$ component of a solution to (DOP) can be discontinuous, the corresponding component z_h in the discretization is allowed to be discontinuous. Theorem 3 then allows one to obtain convergence guarantees even if $z^\star$ is discontinuous. Note that high-order convergence can be guaranteed even if $z^\star$ is discontinuous, provided that it can be approximated in the finite element space, see (4) and the discontinuous elements in Figure 1. It is then a practical matter to employ an adaptive meshing technique for achieving this.

While this work has a theoretical focus, the practicality of our novel transcription has been illustrated in numerical examples. The scheme converges for ill-posed, singular-arc and high-index DAE problems, each of which causes issues for three commonly used direct transcription methods based on collocation, namely TR, HS and LGR. The examples also show that, for the same mesh, the feasibility residual and optimality gap for PBF with $p = 5$ can be orders of magnitude smaller compared to TR and HS, as well as LGR with $p = 5$. This implies that our new method could allow one to achieve the same accuracy as equivalent collocation methods by solving optimization problems that are much smaller in size, but have the same structure as those with collocation methods.

We do not claim in this paper that the best way to compute an approximate solution to (DOP) is by solving the NLP formulated in Section IV. Future work could therefore explore alternative ways of formulating and solving finite-dimensional NLPs, with an in-depth analysis of computational complexity.

The analysis presented here assumes infinite precision. Further work could also include round-off error analysis. Hager notices in a talk[2] that if high-order transcription methods converge, they do so rapidly. Accuracy is then limited by the error in solving the nonlinear program. It is understood that inaccurate solutions to the latter give rise to ringing in the numerical arcs. Thus, caution is

[2]http://users.clas.ufl.edu/hager/Pictures/201010270900-Hager. mp4, minute 37ff.

necessary when interpreting computational results. There is a possibility that LGR would converge to the correct solution in Figure 4 if the NLP could have been solved several orders of magnitude more accurately than can be done using double precision.

Appendix A
Converting a Bolza Problem into (DOP)

Many control, state or parameter estimation problems can be converted into the Bolza form [1], [42] below, possibly over a variable time domain and/or with a terminal cost function f_E (also known as a Mayer term or final cost):

$$\min_{\chi, \upsilon, \xi, \tau_0, \tau_E} \int_{\tau_0}^{\tau_E} f_r \left(\dot{\chi}(\tau), \chi(\tau), \upsilon(\tau), \xi, \tau \right) \mathrm{d}\tau + f_E(\chi(\tau_E), \tau_E)$$

subject to

$$b_B \left(\chi(\tau_0), \chi(\tau_E), \tau_0, \tau_E \right) = 0,$$
$$c_e \left(\dot{\chi}(\tau), \chi(\tau), \upsilon(\tau), \xi, \tau \right) = 0 \text{ f.a.e. } \tau \in (\tau_0, \tau_E),$$
$$c_i \left(\dot{\chi}(\tau), \chi(\tau), \upsilon(\tau), \xi, \tau_0, \tau_E, \tau \right) \leq 0 \text{ f.a.e. } \tau \in (\tau_0, \tau_E),$$

where the state is χ, the weak derivative $\dot{\chi} := \mathrm{d}\chi/\mathrm{d}\tau$ and the input is υ. Note that the starting point τ_0, end point τ_E and a constant vector of parameters $\xi \in \mathbb{R}^{n_\xi}$ are included as optimization variables. The above problem can be converted into the Lagrange form (DOP) as follows.

Move the Mayer term into the integrand by noting that $f_E(\chi(\tau_E), \tau_E) = \phi(\tau_E) = \int_{\tau_0}^{\tau_E} \dot{\phi}(\tau) \mathrm{d}\tau$ if we let $\phi(\tau) := f_E(\chi(\tau), \tau)$ with initial condition $\phi(\tau_0) = 0$. Recall that for minimum-time problems, we usually let $f_E(\chi(\tau), \tau) := \tau$ and $f_r := 0$, so that $\dot{\phi}(\tau) = 1$.

By introducing the auxiliary function s, convert the inequality constraints into the equality constraints $s(\tau) + c_i \left(\dot{\chi}(\tau), \chi(\tau), \upsilon(\tau), \xi, \tau_0, \tau_E, \tau \right) = 0$ and inequality constraint $s \geq 0$. Introduce the auxiliary functions $(\upsilon^+, \upsilon^-) \geq 0$ with the substitution $\upsilon = \upsilon^+ - \upsilon^-$ so that the algebraic variable function is defined as $z := (s, \upsilon^+, \upsilon^-)$.

The problem above with a variable domain is converted onto a fixed domain with $t \in (0, 1)$ via the transformation $\tau = \tau_0 + (\tau_E - \tau_0) \cdot t$ so that $t_0 := 0$, $t_E := 1$.

Define the state for problem (DOP) as $y := (\chi, \phi, \xi, \tau_0, \tau_E)$ and introduce additional equality constraints in order to force $(\mathrm{d}\xi/\mathrm{d}t, \mathrm{d}\tau_0/\mathrm{d}t, \mathrm{d}\tau_E/\mathrm{d}t) = 0$.

The expressions for f, c, b can now be derived using the above.

Appendix B
Lebesgue equivalence for Polynomials

Let $T := (a, b) \in \mathcal{T}_h$ and $p \in \mathbb{N}_0$. We show that

$$\|\beta \cdot u\|_{L^\infty(T)} \leq \frac{p+1}{\sqrt{|T|}} \cdot \|\beta \cdot u\|_{L^2(T)} \quad \forall u \in \mathcal{P}_p(T), \forall \beta \in \mathbb{R}.$$
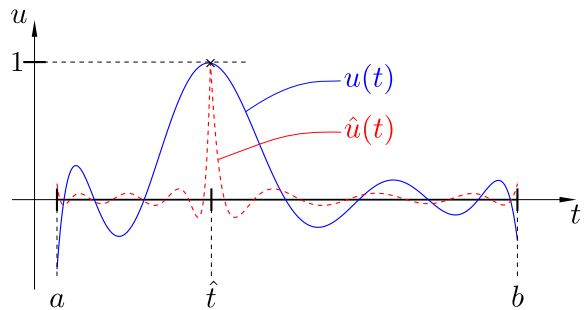

Fig. 10. Polynomial u and piecewise polynomial $\hat{u}$ over T.

Choose $u \in \mathcal{P}_p(T)$ arbitrary. Since $\|\beta \cdot u\|_{L^k(T)} = |\beta| \cdot \|u\|_{L^k(T)}$ holds for both $k \in \{2, \infty\}$, and for all $\beta \in \mathbb{R}$, w.l.o.g. let $\|u\|_{L^\infty(T)} = 1$. Since $\mathrm{sgn}(\beta)$ is arbitrary, w.l.o.g. let $u(\hat{t}) = 1$ for some $\hat{t} \in \overline{T}$. Define $T_L := [a, \hat{t}]$, $T_R := [\hat{t}, b]$, $\hat{\mathcal{P}}_p := \mathcal{P}_p(T_L) \cap \mathcal{P}_p(T_R) \cap \mathcal{C}^0(T)$, and $\hat{u} := \arg\min_{v \in \hat{\mathcal{P}}_p} \{\|v\|_{L^2(T)} \mid v(\hat{t}) = 1\}$. Since $\mathcal{P}_p \subset \hat{\mathcal{P}}_p$, it holds $\|u\|_{L^2(T)} \geq \|\hat{u}\|_{L^2(T)}$. Figure 10 illustrates $u, \hat{u}$ for $p = 8$.

Use $\|\hat{u}\|_{L^2(T)}^2 = \int_a^b \hat{u}(t)^2 \mathrm{d}t = (b-a)/2 \cdot \int_{-1}^1 \hat{u}_{\mathrm{ref}}(t)^2 \mathrm{d}t = \frac{|T|}{2} \cdot \|\hat{u}_{\mathrm{ref}}\|_{L^2(T_{\mathrm{ref}})}^2$, where $\hat{u}_{\mathrm{ref}}$ is $\hat{u}$ linearly transformed from T onto $T_{\mathrm{ref}} := (-1, 1)$. Since $\|\hat{u}\|_{L^2(T)}$ is invariant under changes of $\hat{t}$ because $\hat{u}(\hat{t} + (b - \hat{t}) \cdot \xi) = \hat{u}(\hat{t} + (\hat{t} - a) \cdot \xi)$ $\forall \xi \in [0, 1]$, w.l.o.g. we can assume for $\hat{u}$ that $\hat{t} = b$ and hence $\hat{u}_{\mathrm{ref}}(1) = 1$. Since minimizing the $L^2(T_{\mathrm{ref}})$-norm, $\hat{u}_{\mathrm{ref}}$ solves

$$\min_{v \in \mathcal{P}_p(T_{\mathrm{ref}})} 1/2 \cdot \int_{T_{\mathrm{ref}}} v(t)^2 \mathrm{d}t \text{ subject to } v(1) = 1. \quad (9)$$

We represent $\hat{u}_{\mathrm{ref}} = \sum_{j=0}^p \alpha_j \cdot \phi_j$, where ϕ_j is the jth Legendre polynomial. These satisfy [43]: $\phi_j(1) = 1 \quad \forall j \in \mathbb{N}_0$, $\int_{T_{\mathrm{ref}}} \phi_j(t) \cdot \phi_k(t) \mathrm{d}t = \delta_{j,k} \cdot \gamma_j \quad \forall j, k \in \mathbb{N}_0$, where $\gamma_j := 2/(2 \cdot j + 1)$ and $\delta_{j,k}$ the Kronecker delta. We write $\mathbf{x} = (\alpha_0, \alpha_1, \ldots, \alpha_p)^\mathsf{T} \in \mathbb{R}^{p+1}$, $D = \mathrm{diag}(\gamma_0, \gamma_1, \ldots, \gamma_p) \in \mathbb{R}^{(p+1) \times (p+1)}$ and $\mathbf{1} \in \mathbb{R}^{p+1}$. Then (9) can be written in $\mathbf{x}$:

$$\min_{\mathbf{x} \in \mathbb{R}^{p+1}} \psi(\mathbf{x}) := 1/2 \cdot \mathbf{x}^\mathsf{T} \cdot D \cdot \mathbf{x} \text{ subject to } \mathbf{1}^\mathsf{T} \cdot \mathbf{x} = 1.$$

From the optimality conditions [44, p. 451] follows $\mathbf{x} = D^{-1} \cdot \mathbf{1} \cdot \lambda$ and $\mathbf{1}^\mathsf{T} \cdot D \cdot \mathbf{1} \cdot \lambda = 1$. Using $\mathbf{1}^\mathsf{T} \cdot D \cdot \mathbf{1} = \sum_{j=0}^p \gamma_j = \frac{(p+1)^2}{2}$ yields $\lambda = 1/(p+1)^2$ and $\psi(\mathbf{x}) = \frac{1}{2} \cdot (D^{-1} \cdot \mathbf{1} \cdot \lambda)^\mathsf{T} \cdot D \cdot (D^{-1} \cdot \mathbf{1} \cdot \lambda) = \frac{\lambda}{2} = \frac{1}{(p+1)^2}$. Hence, $\frac{1}{2} \cdot \|\hat{u}_{\mathrm{ref}}\|_{L^2(T_{\mathrm{ref}})}^2 = 1/(p+1)^2$. Hence, $\frac{1}{2} \cdot \|\hat{u}\|_{L^2(T)}^2 = \frac{|T|}{2} \cdot 1/(p+1)^2$. Hence, $\|u\|_{L^2(T)} \geq \|\hat{u}\|_{L^2(T)} = \sqrt{|T|}/(p+1) \cdot \|u\|_{L^\infty(T)}$, or, $\|\beta \cdot u\|_{L^2(T)} \geq \sqrt{|T|}/(p+1) \cdot \|\underbrace{\beta \cdot u}_{\tilde{u}}\|_{L^\infty(T)} \cdot \underbrace{}_{=1}$
In conclusion:

$$\|\tilde{u}\|_{L^\infty(T)} \leq \frac{p+1}{\sqrt{|T|}} \cdot \|\tilde{u}\|_{L^2(T)} \quad \forall \tilde{u} \in \mathcal{P}_p(T) \, \forall T \in \mathcal{T}_h. \quad (10)$$

Appendix C
Order of Approximation for Non-smooth and Smooth Non-differentiable Functions

In the following we illustrate that the assumption $\ell > 0$ in (4) is rather mild. To this end we consider two pathological functions for $g := x^\star_{\omega, \tau}$. In our setting, $n_y = 0$, $n_z = 1$,
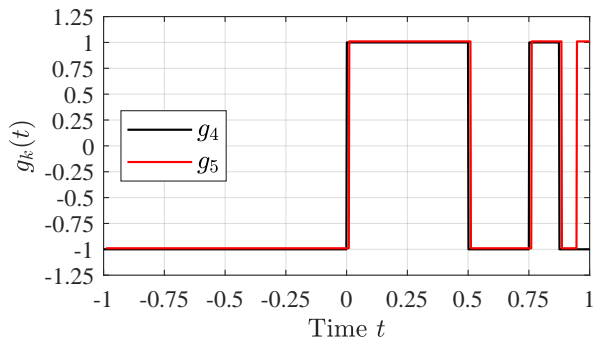
Fig. 11. Nested step-function g_k for $k = 4, 5$.

and we interpolate a given pathological function g with $x_h \in \mathcal{X}_{h,p}$ over $\Omega = (-1, 1)$. We use $p = 0$.

a) A function with infinitely many discontinuities: The first example is a non-smooth function that has infinitely many discontinuities. Similar functions can arise as optimal control solutions. An example is the solution to Fuller's problem [45].

The function under consideration is the limit g_∞ of the following series:

$$g_0(t) := -1,$$

$$g_{k+1}(t) := \begin{cases} g_k(t) & \text{if } t \leq 1 - 2^{-k} \\ -g_k(t) & \text{otherwise} \end{cases} \quad k = 0, 1, 2, \ldots.$$

g_∞ is a function that switches between -1 and 1 whenever t halves its distance to 1. Figure 11 shows g_k for $k = 4, 5$.

Using mesh-size $h = 2^{-k}$ for some $k \in \mathbb{N}$, define $u(t) := g_k(t) \in \mathcal{X}_{h,p}$. Hence,

$$\inf_{x_h \in \mathcal{X}_{h,p}} \{\|g_\infty - x_h\|_{\mathcal{X}}\} \leq \|g_\infty - u\|_{L^2(\Omega)}.$$

It follows that

$$|u(t) - g_\infty(t)| \leq \begin{cases} 0 & \text{if } t \leq 1 - 2^{-k} \\ 2 & \text{otherwise} \end{cases}$$

Hence, $\|g_\infty - u\|_{L^2(\Omega)} \leq \|g_\infty - u\|_{L^1(\Omega)} \leq 2/2^k = \mathcal{O}(h^1)$. Therefore, all $\ell \in (0, 0.5]$ satisfy (4).

b) A continuous but nowhere differentiable function: We consider the following Weierstrass function, which is continuous but non-differentiable:

$$g(t) := \frac{1}{2} \cdot \sum_{k=0}^{\infty} a^k \cdot \cos(7^k \cdot \pi \cdot t)$$

for $0 < a \leq 0.5$. This function with range $\subset [-1, 1]$ satisfies the Hölder property

$$|g(t) - g(s)| \leq C \cdot |t - s|^\alpha$$

with some $C \in \mathbb{R}_+$ for $\alpha = -\log(a)/\log(7)$ [46]. For $a \leq 0.375$ we have $\alpha \geq 0.504$.

According to this property, a piecewise constant interpolation $u \in \mathcal{X}_{h,p}$ of g satisfies

$$|g(t) - u(t)| \leq |g(t) - g(s)| \leq C \cdot |\underbrace{t - s}_{\leq h}|^\alpha.$$

In conclusion,

$$\inf_{x_h \in \mathcal{X}_{h,p}} \{\|g - x_h\|_{\mathcal{X}}\} \leq \|g - u\|_{L^2(\Omega)}$$

$$\leq \|g - u\|_{L^1(\Omega)} = \mathcal{O}(h^\alpha).$$

Therefore, all $\ell \in (0, \alpha - 0.5]$ satisfy (4).

REFERENCES

[1] J. T. Betts, *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming*, 2nd ed. New York, NY, USA: Cambridge University Press, 2010.

[2] B. A. Conway, "A Survey of Methods Available for the Numerical Optimization of Continuous Dynamic Systems," *Journal of Optimization Theory and Applications*, vol. 152, no. 2, pp. 271–306, Feb 2012. [Online]. Available: https://doi.org/10.1007/s10957-011-9918-z

[3] A. V. Rao, "Survey of Numerical Methods for Optimal Control," in *Advances in the Astronautical Sciences*, vol. 135, 2010.

[4] M. A. Patterson and A. V. Rao, "GPOPS-II: A MATLAB software for solving multiple-phase optimal control problems using hp-adaptive Gaussian quadrature collocation methods and sparse nonlinear programming," *ACM Trans. Math. Softw.*, vol. 41, no. 1, pp. 1–37, 2014. [Online]. Available: http://doi.acm.org/10.1145/2558904

[5] M. Kelly, "An introduction to trajectory optimization: How to do your own direct collocation," *SIAM Rev.*, vol. 59, no. 4, pp. 849–904, 2017. [Online]. Available: https://doi.org/10.1137/16M1062569

[6] C. Arévelo, C. Führer, and M. Selva, "A collocation formulation of multistep methods for variable step-size extensions," *Applied Numerical Mathematics*, vol. 42, no. 1–3, pp. 5–16, August 2002.

[7] M. P. Neuenhofen, "High-order convergent finite-elements direct transcription method for constrained optimal control problems," *arXiv:1712.07761*, Dec. 2017.

[8] J. Locker and P. M. Prenter, "Optimal L^2 and L^∞ error estimates for continuous and discrete least squares methods for boundary value problems," *SIAM J. Numer. Anal.*, vol. 15, no. 6, pp. 1151–1160, December 1978.

[9] U. Ascher, "Discrete least squares approximations for ordinary differential equations," *SIAM J. Numer. Anal.*, vol. 15, no. 3, June 1978.

[10] K. Malanowski, C. Büskens, and H. Maurer, "Convergence of approximations to nonlinear optimal control problems," in *Mathematical programming with data perturbations*, ser. Lecture Notes in Pure and Appl. Math. Dekker, New York, 1998, vol. 195, pp. 253–284.

[11] S. Kameswaran and L. T. Biegler, "Simultaneous dynamic optimization strategies: Recent advances and challenges," *Computers and Chemical Engineering*, vol. 30, no. 10, pp. 1560 – 1575, 2006, papers form Chemical Process Control VII. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0098135406001335

[12] G. M. Aly and W. C. Chan, "Application of a modified quasi-linearization technique to totally singular optimal problems," *International Journal of Control*, vol. 17, pp. 809–815, 1973.

[13] S. Kameswaran and L. T. Biegler, "Convergence rates for direct transcription of optimal control problems using collocation at Radau points," *Comput. Optim. Appl.*, vol. 41, pp. 81–126, 2008.

[14] W. W. Hager, J. Liu, S. Mohapatra, A. V. Rao, and X.-S. Wang, "Convergence rate for a Gauss collocation method applied to constrained optimal control," *SIAM J. Control Optim.*, vol. 56, no. 2, pp. 1386–1411, 2018. [Online]. Available: https://doi.org/10.1137/16M1096761

[15] Q. Gong, W. Kang, and I. M. Ross, "A pseudospectral method for the optimal control of constrained feedback linearizable systems," *IEEE Transactions on Automatic Control*, vol. 51, no. 7, pp. 1115–1129, July 2006.

[16] W. Kang, Q. Gong, I. M. Ross, and F. Fahroo, "On the convergence of nonlinear optimal control using pseudospectral methods for feedback linearizable systems," *International Journal of Robust and Nonliner Control*, vol. 17, pp. 1251–1277, 2007.

[17] Q. Gong, I. M. Ross, W. Kang, and F. Fahroo, "Connections between the covector mapping theorem and convergence of pseudospectral methods for optimal control," *Comput. Optim. Appl. Appl*, vol. 41, pp. 307–335, 2008.

[18] M. S. Aronna, J. F. Bonnans, and P. Martinon, "A shooting algorithm for optimal control problems with singular arcs," *J. Optim. Theory Appl.*, vol. 158, no. 2, pp. 419–459, 2013. [Online]. Available: https://doi.org/10.1007/s10957-012-0254-8

[19] G. Pannocchia, J. B. Rawlings, D. Q. Mayne, and W. Marquardt, "On computing solutions to the continuous time constrained linear quadratic regulator," *IEEE Transactions on Automatic Control*, vol. 55, no. 9, pp. 2192–2198, September 2010.

[20] G. Pannocchia, J. B. Rawlings, D. Q. Mayne, and G. M. Mancuso, "Whither discrete time model predictive control?" *IEEE Transactions on Automatic Control*, vol. 60, no. 1, pp. 246–252, January 2015.

[21] J. Yuz, G. Goodwin, and J. De Doná, "Control of constrained linear systems using fast sampling rates," *Systems and Control Letters*, vol. 54, no. 10, pp. 981–990, October 2005.

[22] A. Schwartz and E. Polak, "Consistent approximations for optimal control problems based on Runge-Kutta integration," *SIAM J. Control Optim.*, vol. 34, no. 4, pp. 1235–1269, July 1996.

[23] E. Polak, *Optimization: Algorithms and Consistent Approximations.* Springer-Verlag, 1997.

[24] W. W. Hager, "Runge-Kutta methods in optimal control and the transformed adjoint system," *Numerische Mathematik*, vol. 87, no. 2, pp. 247–282, Dec 2000. [Online]. Available: https://doi.org/10.1007/s002110000178

[25] A. E. Bryson, Jr. and Y. C. Ho, *Applied optimal control.* Hemisphere Publishing Corp. Washington, D. C.; distributed by Halsted Press [John Wiley & Sons], New York-London-Sydney, 1975, optimization, estimation, and control, Revised printing.

[26] R. A. Adams, *Sobolev spaces.* Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975, pure and Applied Mathematics, Vol. 65.

[27] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Math. Program.*, vol. 106, no. 1, Ser. A, pp. 25–57, 2006. [Online]. Available: https://doi.org/10.1007/s10107-004-0559-y

[28] Z. Ding, "A proof of the trace theorem of Sobolev spaces on Lipschitz domains," *Proc. Amer. Math. Soc.*, vol. 124, no. 2, pp. 591–600, 1996. [Online]. Available: https://doi.org/10.1090/S0002-9939-96-03132-2

[29] P. G. Ciarlet, *The finite element method for elliptic problems.* North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978, studies in Mathematics and its Applications, Vol. 4.

[30] J. H. Bramble and S. R. Hilbert, "Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation," *SIAM J. Numer. Anal.*, vol. 7, pp. 112–124, 1970. [Online]. Available: https://doi.org/10.1137/0707006

[31] P. Abbott, "Tricks of the trade: Legendre-gauss quadrature," *Mathematica Journal*, vol. 9, no. 4, pp. 689–691, 2005.

[32] A. Fiacco and G. McCormick, *Nonlinear Programming: Sequential unconstrained minimization techniques.* John Wiley & Sons, 1968.

[33] P. Armand, J. Benoist, R. Omheni, and V. Pateloup, "Study of a primal-dual algorithm for equality constrained minimization," *Comput. Optim. Appl.*, vol. 59, no. 3, pp. 405–433, 2014. [Online]. Available: https://doi.org/10.1007/s10589-014-9679-3

[34] R. Kuhlmann and C. Büskens, "A primal-dual augmented Lagrangian penalty-interior-point filter line search algorithm," *Mathematical Methods of Operations Research*, pp. 1–33, 2017.

[35] Y. Cao, A. Seth, and C. D. Laird, "An augmented Lagrangian interior-point approach for large-scale NLP problems on graphics processing units," *Computers and Chemical Engineering*, vol. 85, pp. 76 – 83, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0098135415003257

[36] A. Forsgren, P. E. Gill, and M. H. Wright, "Interior methods for nonlinear optimization," *SIAM Rev.*, vol. 44, no. 4, pp. 525–597 (2003), 2002. [Online]. Available: https://doi.org/10.1137/S0036144502414942

[37] A. Forsgren and P. E. Gill, "Primal-dual interior methods for nonconvex nonlinear programming," *SIOPT*, vol. 8, no. 4, pp. 1132–1152, 1998. [Online]. Available: https://doi.org/10.1137/S1052623496305560

[38] E. Hairer and G. Wanner, *Solving ordinary differential equations. II*, ser. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2010, vol. 14, stiff and differential-algebraic problems, Second revised edition, paperback. [Online]. Available: https://doi.org/10.1007/978-3-642-05221-7

[39] Y. Nie, O. Faqir, and E. C. Kerrigan, "ICLOCS2: Solve your optimal control problems with less pain," in *Proc. 6th IFAC Conference on Nonlinear Model Predictive Control*, 2018.

[40] H. Maurer, "Theory and applications of bang-bang and singular control problems," 2007, 23rd IFIP TC 7 Conference on System Modelling and Optimization Kraków, Poland, July 23–27.

[41] J. T. Betts, "A collection of optimal control test problems," 2015. [Online]. Available: http://www.appliedmathematicalanalysis.com/downloads/prbook.pdf

[42] J. B. Rawlings, D. Q. Mayne, and M. Diehl, *Model Predictive Control: Theory, Computation, and Design*, 2nd ed. Nob Hill Publishing, LLC, Madison WI, USA, 2017.

[43] E. W. Weisstein. Legendre polynomial. A Wolfram Web Source. [Online]. Available: http://mathworld.wolfram.com/LegendrePolynomial.html

[44] J. Nocedal and S. J. Wright, *Numerical optimization*, 2nd ed., ser. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006.

[45] A. T. Fuller, "Study of an optimum non-linear control system," *J. Electronics Control (1)*, vol. 15, pp. 63–71, 1963.

[46] A. Zygmund, *Trigonometric series. Vol. I, II*, 3rd ed., ser. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2002, with a foreword by Robert A. Fefferman.