Least-squares finite element method for ordinary differential equations

Matthias Chung [a,b,c,*], Justin Krueger [a], Honghu Liu [a]

[a] *Department of Mathematics, Virginia Tech, Blacksburg, VA, United States of America*
[b] *Academy of Data Science, Virginia Tech, Blacksburg, VA, United States of America*
[c] *Department of Mathematics, Emory University, Atlanta, GA, United States of America*

ARTICLE INFO

ABSTRACT

We consider the least-squares finite element method (lsfem) for systems of nonlinear ordinary differential equations, and establish an optimal error estimate for this method when piecewise linear elements are used. The main assumptions are that the vector field is sufficiently smooth and that the local Lipschitz constant as well as the operator norm of the Jacobian matrix associated with the nonlinearity are sufficiently small, when restricted to a suitable neighborhood of the true solution for the considered initial value problem. This theoretic optimality is further illustrated numerically, along with evidence of possible extension to higher-order basis elements. Examples are also presented to show the advantages of lsfem compared with finite difference methods in various scenarios. Suitable modifications for adaptive time-stepping are discussed as well.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

In scientific fields ranging from systems biology and systems engineering to social sciences, physical systems and finance, differential equations are omnipresent and constitute an essential tool to simulate, analyze, predict, and to ultimately make informed decisions. Due to the wide range of applications, the search for efficient, flexible, and reliable numerical schemes is still a timely topic despite its long history. Numerical solutions of ordinary differential equations (ODEs), in particular for initial-value problems (IVPs), are predominantly obtained by a rich variety of finite difference single/multistep schemes, which lead to both implicit and explicit solvers that are now standard in many programming languages [1–5]. In contrast, finite element methods for ODEs are much less investigated, despite the works on continuous and discontinuous Galerkin methods (see [6, Section 2.2] for a brief review) and collocation methods [7]. The same can be said for delay differential equations (DDEs) as well [8].

Motivation. In this work, we initiate an effort to explore the least-squares finite element method (lsfem) as a viable way to numerically solve ODEs and DDEs. Before entering into details, we briefly illustrate the strength of the lsfem using a simple-looking ODE, which turns out to be challenging for traditional finite difference methods (FDMs) to operate. The problem of concern here is the following linear IVP

$$y' = y - 2e^{-t}, \quad y(0) = 1. \tag{1}$$
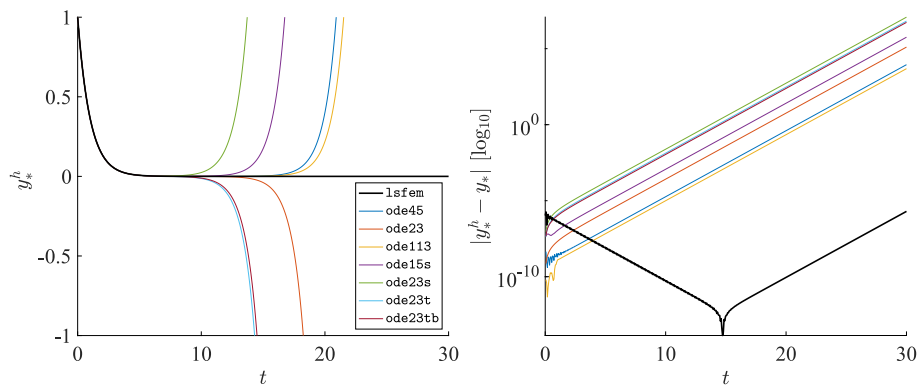
Fig. 1. Failure of standard finite difference methods. Left panel: Numerical solutions of $y' = y - 2e^{-t}$ with $y(0) = 1$ using `lsfem` solver (black bold line) and various standard ODE solvers from Matlab's ODE solver library. The obtained numerical solutions are denoted by y_*^h. Right panel: The point-wise error between the exact solution $y_*(t) = e^{-t}$ and the numerical solutions presented in the left panel as obtained from either the `lsfem` or the various Matlab ODE solvers. For all the Matlab's built-in ODE solvers, the relative and absolute tolerances are set to be `RelTol` $= 10^{-8}$ and `AbsTol` $= 10^{-8}$; and the largest allowed step size is `MaxStep` $= 0.1$. For `lsfem`, we used cubic splines as the basis functions defined on a uniform mesh of size $\delta t = 0.1$.

Note that the exact solution is given by $y(t) = e^{-t}$. Remarkably, all Matlab built-in numerical ODE solvers fail on this example when solution over a relatively long time interval is computed. Whereas, the proposed `lsfem` tracks well the exact solution. In Fig. 1, we present the numerical solutions (left panel) and the corresponding pointwise errors (right panel) on the interval $t \in [0, 30]$ for all these solvers. As can be seen in the left panel, sooner or later, the solutions from the finite-difference schemes exhibit exponential growth, leading thus to exponentially growing pointwise errors. In contrast, the maximum error for `lsfem` over the whole interval remains below $2 \cdot 10^{-6}$ (see black curve in the right panel of Fig. 1). It is also worth noting that the setup of the experiment is actually in favor of the built-in solvers, since we used a uniform mesh size for `lsfem` while allowing the build-in Matlab solvers to exhibit smaller or equal step sizes compared with the mesh size for `lsfem`; see the caption of Fig. 1 for further details.

The failure of the FDMs for the above example is actually not surprising. It results from discretization errors which are amplified exponentially over time since the equation has no stabilizing nonlinear terms to counterbalance the linear instability. Indeed, assume that at a given time instant $s > 0$, the true solution $y(s) = e^{-s}$ is perturbed by a small amount ϵ, that is $\widehat{y}(s) = e^{-s} + \epsilon$. Then, by direct calculation using the original equation, one sees that this deviation gets amplified to $e^{t-s}\epsilon$ for all $t \geq s$. Since local discretization errors are intrinsic to any FDMs, such deviations are unavoidable.

In contrast to the "localization" nature of FDMs, the aim of an `lsfem` is to find an optimal approximate solution within a given subspace that minimizes an objective function over the whole time interval of integration (cf. Section 2.1), hence making such methods much more robust to local discretization errors compared to FDMs. The `lsfem` methods are also flexible in the sense that minor changes are needed when considering different types of dynamical systems, either governed by ODEs or DDEs and in the contexts of either IVPs or boundary value problems (BVPs), which allows for a unified numerical implementation for all the cases. In fact, the setup can also handle a broad class of differential algebraic equations (DAEs) as well, with the associated optimization problems become now constrained optimizations. Moreover, since the objective function directly controls the discretization error, it can be used as a diagnostic tool for local mesh adaptivity consideration, a feature crucial for problems involving abrupt local changes or stiffness.

Since their emergence in the early 1950s, finite element methods (FEMs) have become one of the most versatile and powerful methodologies for the numerical solution of partial differential equations (PDEs). Whereas, for ODEs and DDEs, the usage of FEMs is much less pursued as mentioned above. Intuitively, this may be related to the facts that the salient feature of geometrical flexibility of FEMs is dormant in these cases, and that solutions for ODEs and DDEs are oftentimes smooth, rendering the weak formulation of FEMs less attractive.

However, as already illustrated in Fig. 1, the `lsfem` can provide accurate solutions in situations that traditional FDMs may fail drastically. This is further supported by other examples in Section 4 that the superior performance of the `lsfem` reported in Fig. 1 is not just an exception. These numerical results prompt us to re-evaluate the aforementioned intuition about the usage of FEMs, at least in the least-squares settings, for ODEs and DDEs.

These investigations are further driven by newly discovered connections between ordinary differential equations and residual neural networks [9]. Within the field of neural networks where stability is a major concern, recent works are starting to investigate finite element type solvers [10].

The existing literature on `lsfem` is mainly devoted to PDEs; see e.g., [11–14] and references therein. On the theoretic side, for linear PDE problems, very satisfactory theoretical understandings have already been gained that includes convergence results and even optimal error estimates [12,14]. Nevertheless, error estimates in the case of nonlinear PDE problems remain largely open. In contrast, `lsfem` for ODEs and DDEs has not received much attention yet, neither theoretically nor computationally. In this article, we take a first step in establishing `lsfem` error analysis for general nonlinear ODEs, deferring the treatments for DAEs and DDEs to future works.

Main contributions. In that respect, we consider IVP of nonlinear ODEs for which we establish under suitable conditions optimal error estimates for `lsfem` with piecewise linear elements; see Theorem 3.1. The optimality of the estimate is in the sense that the error bound established in Theorem 3.1 is of the same order, in terms of the mesh size h, as the finite element interpolation error recalled in Lemma 3.2. Our main idea centers around an estimate given by Proposition 3.2 associated with an auxiliary system (23). Given an `lsfem` solution y_*^h in a finite dimensional subspace X^h of the Sobolev (state) space $X = H^1(0, T; \mathbb{R}^d)$, this latter auxiliary system is obtained by replacing the original nonlinear vector field $F(y) + f(t)$ in (3) by $F(y_*^h(t)) + f(t)$. Since the latter vector field consists simply of a given time-dependent function for a given y_*^h, optimal error estimates between the true solution w_* and the `lsfem` solution w_*^h of the auxiliary system (23) is well known using the classical Aubin–Nitsche trick [15–17]; see Lemma 3.1 and the estimate given by (25), in which the dependence on y_*^h are marked via $w_*[y_*^h]$ and $w_*^h[y_*^h]$. However, to establish a suitable control of the difference $\|y_*^h - w_*^h[y_*^h]\|$ between the `lsfem` solutions y_*^h (for the original nonlinear IVP (3)) and $w_*^h[y_*^h]$ (for the auxiliary system (23)) as given in Proposition 3.2 requires a major effort.

Such an estimate for $\|y_*^h - w_*^h[y_*^h]\|$ is established through a series of lemmas that exploit geometric properties revealed by the first-order optimality condition associated with each minimizer y_*^h in the subspace X^h for the objective function J given by (4). Indeed, from $\frac{d}{d\tau}J(y_*^h + \tau v_h; F, f, g)\big|_{\tau=0} = 0$ for all v_h in X^h, after some algebraic operations, we can actually link this necessary condition with $w^*[y_*^h]$ through the following orthogonality property (cf. Lemma 3.3):

$$\langle y_*^h - w_*[y_*^h], \ v_h - \Gamma(\cdot; v_h)\rangle_X = 0, \quad \forall \ v_h \in X^h, \tag{2}$$

where Γ is an integral involving the Jacobian matrix of F given by (29). It is this simple, albeit not so obvious, geometric identity that opens the room for estimation, once Γ is further split as the sum of its projection $\Pi_h \Gamma$ onto X^h and its orthogonal complement $\Pi_h^\perp \Gamma$; see Lemmas 3.4 and 3.5.

Although the error analysis presented in this article focuses on piecewise linear elements, numerical evidence provided in Section 4 indicates that when a piecewise spline basis of degree k is used to form X^h and F is C^{k+1}-smooth, then the error bound scales like h^{k+1}. Rigorous justification of such an error estimate will be addressed in a future work.

Organization. This article is organized as follows. We first recall in Section 2 the basic setup of `lsfem` in the context of IVP for nonlinear ODE systems. Besides its functional framework recalled in Section 2.1, for later usage we also present in Section 2.2 a result concerning the convergence of `lsfem` solutions to the true solution; see Theorem 2.1. While the treatment makes a direct usage of a general convergence result on the approximation of abstract nonlinear equations (cf. [18, Theorem 3.3, p.307] and [12, Theorem 8.1]) some detailed calculation is required to recast the problem into the functional form dealt with in [18, Theorem 3.3, p.307] and also to check the required assumptions therein. We provide thus a proof of this convergence theorem in Appendix A for the sake of clarity. The associated optimal error analysis reviewed above is then dealt with in Section 3. The algorithmic aspects are then presented in Section 4 (cf. Algorithm 1) along with numerical results for various concrete examples that confirm the error bounds obtained in Section 3 and also provide numerical evidence for possible extension to higher-order basis elements. We also discuss within this section suitable modifications for adaptive time stepping. Finally, Section 5 provides a brief conclusion and potential future directions.

2. Preliminaries

As a preparation for later sections concerning the error estimates (Section 3) as well as the numerical treatments (Section 4), we briefly summarize the basic setup for `lsfem` of first-order ODEs and then recall a classical convergence result for the `lsfem` solutions. For ease of reference, a table of the main symbols used in this work is provided in Table 1.

2.1. Formulation of `lsfem`

We provide in this subsection a brief account of the `lsfem` for first-order (nonlinear) ODE systems; and refer to [14, Chap. 3] for more details. Given a fixed $T > 0$, consider the following initial-value problem (IVP) in $\mathbb{R}^d$ for some $d \in \mathbb{N}$:

$$\begin{aligned} y' &= F(y) + f(t), \quad t \in (0, T], \\ y(0) &= g, \end{aligned} \tag{3}$$

where $F: \mathbb{R}^d \to \mathbb{R}^d$ is a given smooth and possibly nonlinear function, f is a function in $L^2(0, T; \mathbb{R}^d)$, and g is a given vector in $\mathbb{R}^d$. Precise smoothness on F will be specified later on, and additional regularity on f will be added when optimal error estimates are considered in Section 3.

Before proceeding, it is worth mentioning that all the results of both the current section and Section 3 hold for more general systems of the form $y' = F(t, y) + f(t)$ as well. Indeed, by introducing an auxiliary scalar equation $p' = 1$ supplemented with $p(0) = 0$, and considering the new variable $z = (p, y)^\top$, we get $z' = \tilde{F}(z) + \tilde{f}(t)$ with $\tilde{F}(z) = [1, F(p, y)]^\top$ and $\tilde{f} = [0, f]^\top$. This latter system for z is an equivalent formulation of the original problem and fits into the form given by (3).

Throughout the article, we denote the classical Sobolev space $H^1(0, T; \mathbb{R}^d)$ by X, which consists of $L^2(0, T; \mathbb{R}^d)$ functions whose first-order weak derivative is also in $L^2(0, T; \mathbb{R}^d)$. X will be equipped with a norm that is equivalent to the usual

Table 1

List of main symbols.

X	The Sobolev space $H^1(0, T; \mathbb{R}^d)$ equipped with the inner product (7) and the corresponding induced norm (8)
X^h	A finite element subspace of X
I_h	The interpolation operator from X to X^h
Π_h	The orthogonal projection from X to X^h
Id_X	The identity map on X
$\Pi_h^\perp$	Orthogonal complement of Π_h: $\Pi_h^\perp = \mathrm{Id}_X - \Pi_h$
y_*	Solution to the variational formulation (5) of the IVP (3)
y_*^h	`lsfem` approximation of y_* in the subspace X^h; i.e., solution of (6)
$w_*[y_*^h]$	Solution of the auxiliary system (23)
$w_*^h[y_*^h]$	`lsfem` approximation of $w_*[y_*^h]$ in the subspace X^h
u	A generic element in X or the solution of (A.5) depending on the context
v	A generic element in X
$\mathfrak{B}$	The subset in $\mathbb{R}^d$ defined by (14), which contains both $y_*(t)$ and the `lsfem` solution $y_*^h(t)$ for all t in $[0, T]$ and all sufficiently small h
$\mathfrak{C}$	Embedding constant for the continuous embedding from X to $C([0, T]; \mathbb{R}^d)$
$\tilde{\mathfrak{C}}$	Embedding constant for the continuous embedding from X to $L^2(0, T; \mathbb{R}^d)$
$\langle \cdot, \cdot \rangle$	The standard dot product on $\mathbb{R}^d$
$\| \cdot \|$	The Euclidean norm on $\mathbb{R}^d$
$\| \cdot \|_{\mathrm{op}}$	The operator norm for a $d \times d$ matrix, i.e., $\|M\|_{\mathrm{op}} = \sup_{z \in \mathbb{R}^d, \|z\|=1} \|Mz\|$
$L(Y, Z)$	The set of bounded linear maps from a Hilbert space Y to a Hilbert space Z

H^1-norm; see (8). Recall that a function $y \in X$ is called a strong solution of (3) if $y(0) = g$, and $y' = F(y) + f(t)$ for almost every t in $(0, T)$.

The `lsfem` for the IVP (3) relies on a variational reformulation of the ODE system, which seeks for $y_* \in X$ that minimizes the following objective function

$$J(y; F, f, g) := \tfrac{1}{2} \|y' - F(y) - f\|_{L^2(0,T;\mathbb{R}^d)}^2 + \tfrac{1}{2} \|y(0) - g\|^2, \tag{4}$$

where $\| \cdot \|$ denotes the Euclidean norm on $\mathbb{R}^d$. Note that if the IVP (3) admits a unique strong solution in X, then this solution is also the unique solution of the following unconstrained minimization problem:

$$\text{Find} \quad \arg\min_{y \in X} J(y; F, f, g). \tag{5}$$

Given any finite element subspace X^h of X, with h denoting the maximal length of the finite elements, the `lsfem` for the IVP (3) consists of solving the following analogue of the unconstrained minimization problem (5) restricted to X^h:

$$\text{Find} \quad \arg\min_{y^h \in X^h} J(y^h; F, f, g). \tag{6}$$

Let us introduce the following inner product on X, which is naturally related to the objective function J defined in (4), i.e.,

$$\langle u, v \rangle_X = \int_0^T \langle u'(t), v'(t) \rangle \, \mathrm{d}t + \langle u(0), v(0) \rangle, \quad \forall\, u, v \in X, \tag{7}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product in $\mathbb{R}^d$. The norm on X induced by the above inner product $\langle \cdot, \cdot \rangle_X$ will be denoted by $\| \cdot \|_X$, which is often referred to as the energy norm in the literature, namely,

$$\|u\|_X = \sqrt{\langle u, u \rangle_X} = \left(\int_0^T \langle u'(t), u'(t) \rangle \, \mathrm{d}t + \langle u(0), u(0) \rangle \right)^{1/2}, \quad u \in X. \tag{8}$$

One can check by using basic Sobolev inequalities that the $\| \cdot \|_X$-norm is equivalent to the usual Sobolev norm on $H^1(0, T; \mathbb{R}^d)$ defined by $\|u\|_{H^1} = \left(\|u\|_{L^2(0,T;\mathbb{R}^d)}^2 + \|u'\|_{L^2(0,T;\mathbb{R}^d)}^2 \right)^{1/2}$. Note, there exist positive constants c_1 and c_2 such that for all $u \in X$ it holds that $c_1 \|u\|_X \leq \|u\|_{H^1} \leq c_2 \|u\|_X$.

For later usage, let us also introduce two embedding constants. First note that since $H^1(0, T; \mathbb{R}^d)$ is continuously embedded into $C([0, T]; \mathbb{R}^d)$, see e.g., [19, Theorem 8.8], then X equipped with the norm defined in (8) is also continuously embedded into $C([0, T]; \mathbb{R}^d)$. Throughout this article, we denote by $\mathfrak{C}$ the associated embedding constant, where $\mathfrak{C}$ is the smallest constant such that[1]

$$\max_{t \in [0,T]} \|u(t)\| \leq \mathfrak{C} \|u\|_X, \quad \forall\, u \in X. \tag{9}$$

---

[1] For each $u \in X$, we always consider its continuous representative in the corresponding equivalent class. There exists a unique such representative for each $u \in X$; cf. [19, Theorem 8.2].

We denote also by $\widetilde{\mathfrak{c}}$ the embedding constant for the continuous embedding from X to $L^2(0, T; \mathbb{R}^d)$, which is the smallest constant such that

$$\|u\|_{L^2(0,T;\mathbb{R}^d)} \le \widetilde{\mathfrak{c}}\|u\|_X, \quad \forall u \in X. \tag{10}$$

2.2. Convergence of the `lsfem` *solutions*

To prepare for the error analysis carried out in Section 3, we summarize in this subsection a convergence theorem for the `lsfem` solutions as the dimension of the subspace X^h in (6) increases. The treatment makes a direct use of a general result on approximation of abstract nonlinear equations; cf. [18, Theorem 3.3, p.307] and [12, Theorem 8.1].

We work with a sequence of finite element subspaces $\{X^h \subset X\}$, with h denoting the maximal length of the finite elements, such that

$$\lim_{h \to 0} \|(\mathrm{Id}_X - \Pi_h)v\|_X = 0, \quad \forall v \in X, \tag{11}$$

where $\Pi_h : X \to X^h$ denotes the orthogonal projection onto X^h under the inner product $\langle \cdot, \cdot \rangle_X$ defined in (7).

We denote by $\mathrm{D}F$ the Jacobian matrix of F, and by $\|\cdot\|_{\mathrm{op}}$ the operator norm of a bounded linear map from $\mathbb{R}^d$ onto itself.

Theorem 2.1. *Consider the IVP* (3). *Assume that* $f \in L^2(0, T; \mathbb{R}^d)$, $F : \mathbb{R}^d \to \mathbb{R}^d$ *is* C^3 *smooth, and* (3) *has a unique strong solution* y_* *in* X. *Assume also that* $\|\mathrm{D}F(y_*(t))\|_{\mathrm{op}}$ *is sufficiently small for all* $t \in [0, T]$. *Let* $\mathcal{O}$ *be any given open neighborhood of* y_* *in* X, *and* $\{X^h \subset X\}$ *be a sequence of finite element subspaces satisfying* (11). *Then problem* (6) *has a unique solution* y_*^h *in* $\mathcal{O}$ *for all sufficiently small* h, *and* y_*^h *converges in* X-*norm to the solution* y_* *of* (5) *as* h *is reduced to zero,*

$$\lim_{h \to 0} \|y_* - y_*^h\|_X = 0. \tag{12}$$

Since some detailed calculation is required to recast the problem into the functional form dealt with in [18, Theorem 3.3, p.307] and also to check the required assumptions therein, we provide a proof of the above theorem in Appendix A for the sake of clarity.

With the above convergence result available, we are ready to address the associated error analysis. In particular, we show for the case of piecewise linear elements that the `lsfem` achieve optimal rate of convergence, which is the rate dictated by the interpolation error.

3. Optimal `lsfem` error estimates for nonlinear ODEs

In this section, we derive an optimal error estimates for the `lsfem` solutions for first-order nonlinear ODE system of the form (3). The results are obtained for piecewise linear finite elements. Under suitable assumptions, it is shown that the error bound for `lsfem` solutions is proportional to the square of the mesh size, which is of the same order as the interpolation error for piecewise linear finite elements.

Let us first introduce the following assumption about the IVP (3):

(A1) $f : [0, T] \to \mathbb{R}^d$ is absolutely continuous, f' belongs to $L^2(0, T; \mathbb{R}^d)$, and $F : \mathbb{R}^d \to \mathbb{R}^d$ is C^3 smooth. The IVP (3) has a unique solution y_* in X.

Except the strengthened smoothness and integrability requirements on f, the other parts in Assumption **(A1)** are the same as those required in Theorem 2.1.

In Theorem 2.1, a smallness assumption is also made on, $\|\mathrm{D}F(y_*(t))\|_{\mathrm{op}}$, the operator norm of the Jacobian matrix $\mathrm{D}F$ along the solution trajectory y_*. For the derivation of error estimates, this technical assumption needs to be further strengthened and augmented to require that both $\|\mathrm{D}F\|_{\mathrm{op}}$ and the local Lipschitz constant of F are sufficiently small over a bounded set in $\mathbb{R}^d$ that contains the solution y_* as well as the `lsfem` solutions for all time $t \in [0, T]$.

We make precise these smallness assumptions on F below for the sake of clarity. Let us first note that the smallness of $\|\mathrm{D}F(y_*(t))\|_{\mathrm{op}}$ required in Theorem 2.1 is made precise in its proof given by Appendix A. It suffices to require that (see (A.24))

$$\sup_{t \in [0,T]} \|\mathrm{D}F(y_*(t))\|_{\mathrm{op}} < \frac{1}{\sqrt{2T^2 + T + \widetilde{\mathfrak{c}}^2 + \sqrt{(2T^2 + T + \widetilde{\mathfrak{c}}^2)^2 + 2T\widetilde{\mathfrak{c}}^2(1 + 2T)}}}, \tag{13}$$

where $\widetilde{\mathfrak{c}}$ denotes the embedding constant for the continuous embedding from X to $L^2(0, T; \mathbb{R}^d)$; cf. (10).

To present the needed augmentations of (13), we first establish some notations which will be used throughout this section. We take the neighborhood $\mathcal{O}$ of y_* in Theorem 2.1 to be an open ball in X centered at y_* with some radius $r > 0$, which is denoted by $B(y_*, r)$. Let $\bar{h} > 0$ be chosen such that for each $h \in (0, \bar{h})$, the `lsfem` problem (6) has a unique solution y_*^h in $B(y_*, r)$; the existence of such an $\bar{h}$ is guaranteed by Theorem 2.1.

With the embedding constant $\mathfrak{C}$ that ensures (9), we define then

$$\mathfrak{B} = \bigcup_{t \in [0,T]} \left\{ p \in \mathbb{R}^d \ : \ \|p - y_*(t)\| < \mathfrak{C}r \right\}. \tag{14}$$

Since y_*^h stays in $B(y_*, r) \subset X$ for all $h \in (0, \overline{h})$, it holds that $\|y_*^h(t) - y_*(t)\| \le \mathfrak{C}\|y_*^h - y_*\|_X < \mathfrak{C}r$, namely,

$$y_*^h(t) \in \mathfrak{B}, \quad \forall\, h \in (0, \overline{h}),\ t \in [0, T]. \tag{15}$$

The aforementioned smallness assumptions on F are as follows:

(A2) Assume that (13) holds. Let $r > 0$ be arbitrarily given and $\overline{h} > 0$ be chosen so that the `lsfem` solution $y_*^h \in X^h$ stays in the ball $B(y_*, r) \subset X$ for all $h \in (0, \overline{h})$. Let $\mathfrak{B}$ be the subset in $\mathbb{R}^d$ defined by (14) that contains both $y_*(t)$ and the `lsfem` solutions $y_*^h(t)$ for all $h \in (0, \overline{h})$ and all $t \in [0, T]$; cf. (15). Assume that the local Lipschitz constant of F over $\mathfrak{B}$ satisfies

$$\frac{\mathrm{Lip}(F|_{\mathfrak{B}})T}{\sqrt{2}} < 1, \tag{16}$$

and that its Jacobian matrix satisfies

$$\sup_{z \in \mathfrak{B}} \|DF(z)\|_{\mathrm{op}} < \frac{1}{\widetilde{\mathfrak{C}}}, \tag{17}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm of a bounded linear map from $\mathbb{R}^d$ onto itself, and $\widetilde{\mathfrak{C}}$ is the same as given in (13).

Of course, all the three conditions (13), (16), and (17) in **(A2)** can be summarized into one assumption of the form $\sup_{z \in \mathfrak{B}} \|DF(z)\|_{\mathrm{op}} < C$ with C taken to be the right-hand-side (RHS) of (13). However, we prefer to keep them separate in the hope for future improvements since they are used in separate parts of the proof.

The main result of this section is summarized in the following theorem.

Theorem 3.1. *Given a sequence of subspaces $\{X^h \subset X\}$ satisfying (11) and spanned by piecewise linear basis functions, let us consider for each X^h the least-squares finite element approximation (6) of the nonlinear IVP (3). Assume the assumptions **(A1)** and **(A2)** hold. Let $\overline{h}$ be as given in **(A2)**. Then, there exists a constant $C > 0$ independent of h such that the `lsfem` solution y_*^h of (3) satisfies:*

$$\|y_*^h - y_*\|_{L^2(0,T;\mathbb{R}^d)} \le Ch^2, \qquad \forall\, h \in (0, \overline{h}). \tag{18}$$

We first recall a well known error estimate for the special case of the IVP (3), in which F is identically zero. We consider for the moment

$$\begin{aligned}
\widetilde{y}' &= f(t), \quad t \in (0, T], \\
\widetilde{y}(0) &= g.
\end{aligned} \tag{19}$$

In this case, its solution is obviously given by

$$\widetilde{y}_*(t) = g + \int_0^t f(s)\, \mathrm{d}s, \quad t \in [0, T]. \tag{20}$$

For a given finite element subspace X^h, the corresponding `lsfem` approximation $\widetilde{y}_*^h$ is obtained by solving

$$\operatorname*{arg\,min}_{\widetilde{y}^h \in X^h} \ \tfrac{1}{2}\|(\widetilde{y}^h)' - f\|_{L^2(0,T;\mathbb{R}^d)}^2 + \tfrac{1}{2}\|\widetilde{y}^h(0) - g\|^2. \tag{21}$$

Lemma 3.1. *Consider the problem (19). Assume that $f : [0, T] \to \mathbb{R}^d$ is absolutely continuous and f' belongs to $L^2(0, T; \mathbb{R}^d)$. Assume also that X^h is spanned by piecewise linear basis functions. Then, for each such X^h, there exists a unique `lsfem` solution that solves (21), which is given by $\widetilde{y}_*^h = \Pi_h \widetilde{y}_*$, where $\widetilde{y}_*$ is the solution of (19) and Π_h denotes the orthogonal projection from X onto X^h. Moreover, there exists a positive constant C independent of h such that*

$$\|\widetilde{y}_* - \widetilde{y}_*^h\|_{L^2(0,T;\mathbb{R}^d)} \le Ch^2 \|\widetilde{y}_*''\|_{L^2(0,T;\mathbb{R}^d)}. \tag{22}$$

Although the above results are classical, we provide in Appendix B some elements of the proof for the sake of completeness.

Note that since the L^2-error of piecewise linear interpolation for general function in $H^2(0, T; \mathbb{R}^d)$ is of the order h^2 (cf. Lemma 3.2), the above result shows that the corresponding `lsfem` provides the optimal convergence rate for the special case (19). However, the proof of Lemma 3.1 admits no straightforward extension to the general nonlinear case.

To bridge the gap between the setting of Lemma 3.1 (dealing with $F = 0$) and that of Theorem 3.1 (dealing with general nonlinear F), we introduce now an auxiliary system that will serve as a pivot in the estimates presented below. We consider, for a given lsfem solution y_*^h of the IVP (3), the following auxiliary system

$$
\begin{aligned}
w' &= F(y_*^h(t)) + f(t), \quad t \in (0, T], \\
w(0) &= g.
\end{aligned}
\tag{23}
$$

First note that the IVP (23) fits into the form of (19) since $F(y_*^h(t)) + f(t)$ is known once y_*^h is given. Lemma 3.1 is thus applicable. It follows that (23) always admits a unique lsfem solution $w_*^h[y_*^h]$ under the assumption of Lemma 3.1. Moreover, denoting by $w_*[y_*^h]$ the solution of (23), it holds that

$$
w_*^h[y_*^h] = \Pi_h w_*[y_*^h],
\tag{24}
$$

and that

$$
\|w_*[y_*^h] - w_*^h[y_*^h]\|_{L^2(0,T;\mathbb{R}^d)} \le Ch^2 \|(w_*[y_*^h])''\|_{L^2(0,T;\mathbb{R}^d)}.
\tag{25}
$$

We aim to derive the following estimate of $\|y_*^h - w_*^h[y_*^h]\|_{L^2(0,T;\mathbb{R}^d)}$.

Proposition 3.2. *For each given X^h, the following estimate holds for the lsfem solution y_*^h of the nonlinear problem (3) and the lsfem solution $w_*^h[y_*^h]$ of the problem (23):*

$$
\|y_*^h - w_*^h[y_*^h]\|_X \le \frac{C\|(w_*[y_*^h])''\|_{L^2(0,T;\mathbb{R}^d)}}{1 - \widetilde{\mathfrak{C}} \sup_{z \in \mathfrak{B}} \|DF(z)\|_{op}} h^2,
\tag{26}
$$

where $C > 0$ denotes a universal constant independent of h, and $\widetilde{\mathfrak{C}}$ denotes the embedding constant between $L^2(0, T; \mathbb{R}^d)$ and X.

We present next a few lemmas that will be used in the proof of the above Proposition.

Lemma 3.2. *Let X^h be a subspace of X spanned by piecewise linear basis functions. Given any $f \in X$, denote by $I_h f$ the interpolant of f in X^h. Then, there exists a positive constant C independent of h such that the following inequalities hold for all f in the subspace $H^2(0, T; \mathbb{R}^d) \subset X$:*

$$
\|f - I_h f\|_{L^2(0,T;\mathbb{R}^d)} \le Ch^2 \|f''\|_{L^2(0,T;\mathbb{R}^d)},
\tag{27a}
$$

$$
\|f' - (I_h f)'\|_{L^2(0,T;\mathbb{R}^d)} \le Ch \|f''\|_{L^2(0,T;\mathbb{R}^d)},
\tag{27b}
$$

$$
\|f - \Pi_h f\|_X \le Ch \|f''\|_{L^2(0,T;\mathbb{R}^d)}.
\tag{27c}
$$

The first two inequalities above are classical; see e.g., [14, Section 2.5] for a proof. The estimate (27c) follows from (27b) by noting that

$$
\|f - \Pi_h f\|_X \le \|f - I_h f\|_X = \|f' - (I_h f)'\|_{L^2(0,T;\mathbb{R}^d)}.
\tag{28}
$$

The first inequality in (28) holds because for any $f \in X$, its projection $\Pi_h f$ minimizes the residual error $\|f - v^h\|_X$ among all $v^h \in X^h$. Note also that since $t = 0$ is an interpolation point of the piecewise linear finite element subspace X^h, it holds that $f(0) - I_h f(0) = 0$. The second equality in (28) follows.

Lemma 3.3. *Let y_*^h be the lsfem solution of the nonlinear problem (3) as specified in Theorem 3.1. Let $w_*[y_*^h]$ be the solution of the auxiliary problem (23). We define*

$$
\Gamma(t; v^h) = \int_0^t DF(y_*^h(s)) v^h(s) \, ds, \qquad t \in [0, T], \ v^h \in X^h.
\tag{29}
$$

Then, the following identity holds

$$
\langle y_*^h - w_*[y_*^h], \ v^h - \Gamma(\cdot; v^h) \rangle_X = 0, \quad \forall \, v^h \in X^h.
\tag{30}
$$

Proof. The equality (30) is just a reformulation of the first-order necessary condition for y_*^h to be a solution of the minimization problem (6). Indeed, note that this latter condition is given by

$$
\int_0^T \langle (y_*^h)' - F(y_*^h) - f, (v^h)' - DF(y_*^h) v^h \rangle \, dt + \langle y_*^h(0) - g, v^h(0) \rangle = 0, \ \forall \, v^h \in X^h,
\tag{31}
$$

see (A.14) in Appendix A. Note also that

$$
w_*[y_*^h] = g + \int_0^t F(y_*^h(s)) + f(s) \, ds.
$$

Then, (30) follows from (31) by simply noting that $(w_*[y_*^h])'(t) = F(y_*^h(t)) + f(t)$, $\Gamma'(t; v^h) = DF(y_*^h(t)) v^h(t)$, $w_*[y_*^h](0) = g$, and $\Gamma(0; v^h) = 0$. $\square$

The above identity (30) serves as the starting point of our estimates for the term $y_*^h - w_*^h[y_*^h]$. For this purpose, we split Γ defined by (29) as

$$\Gamma(t; v^h) = \Pi_h \Gamma(t; v^h) + \Pi_h^\perp \Gamma(t; v^h), \tag{32}$$

where $\Pi_h^\perp = \mathrm{Id}_X - \Pi_h$.

To simplify the notations, we also denote

$$\gamma = y_*^h - w_*[y_*^h]. \tag{33}$$

Using (32) and (33) in (30), we obtain

$$\langle \gamma, v^h - \Pi_h \Gamma(\,\cdot\,; v^h) \rangle_X = \langle \gamma, \Pi_h^\perp \Gamma(\,\cdot\,; v^h) \rangle_X = \langle \Pi_h^\perp \gamma, \Pi_h^\perp \Gamma(\,\cdot\,; v^h) \rangle_X, \quad \forall\, v^h \in X^h. \tag{34}$$

The estimation of the RHS in the above identity will be considered in Lemma 3.4; and the left-hand-side (LHS) will be considered in Lemma 3.5.

Lemma 3.4. *Let Γ and γ be defined in (29) and (33), respectively. Let $\bar{h}$ be as specified in Theorem 3.1. Then, there exists a constant $C > 0$ independent of h, such that for any $h \in (0, \bar{h})$, it holds that*

$$|\langle \Pi_h^\perp \gamma, \Pi_h^\perp \Gamma(\,\cdot\,; v^h) \rangle_X| \le C \|(w_*[y_*^h])''\|_{L^2(0,T;\mathbb{R}^d)} \|v^h\|_X h^2, \quad \forall\, v^h \in X^h. \tag{35}$$

Proof. The result follows essentially from the estimate (27c) in Lemma 3.2. First note that since $\gamma = y_*^h - w_*[y_*^h]$ and $y_*^h \in X^h$, we have $\Pi_h^\perp \gamma = \Pi_h^\perp(y_*^h - w_*[y_*^h]) = -\Pi_h^\perp w_*[y_*^h]$. This together with (27c) implies that

$$\|\Pi_h^\perp \gamma\|_X \le Ch \|(w_*[y_*^h])''\|_{L^2(0,T;\mathbb{R}^d)}. \tag{36}$$

Again by (27c), we have also

$$\|\Pi_h^\perp \Gamma(\,\cdot\,; v^h)\|_X \le Ch \|\Gamma''(\,\cdot\,; v^h)\|_{L^2(0,T;\mathbb{R}^d)}. \tag{37}$$

It remains to estimate $\|\Gamma''(\,\cdot\,; v^h)\|_{L^2(0,T;\mathbb{R}^d)}$.

Since $\Gamma'(t; v^h) = \mathrm{D}F(y_*^h(t))v^h(t)$, we get for almost every t in $[0, T]$ that

$$\Gamma''(t; v^h) = [\mathrm{D}^2 F(y_*^h(t))(y_*^h(t))']v^h(t) + \mathrm{D}F(y_*^h(t))(v^h(t))'.$$

Then,

$$\|\Gamma''(\,\cdot\,; v^h)\|_{L^2(0,T;\mathbb{R}^d)}^2 = \int_0^T \left\| [\mathrm{D}^2 F(y_*^h(t))(y_*^h(t))']v^h(t) + \mathrm{D}F(y_*^h(t))(v^h(t))' \right\|^2 \, \mathrm{d}t. \tag{38}$$

Note that

$$\begin{aligned}
\int_0^T \left\| [\mathrm{D}^2 F(y_*^h(t))(y_*^h(t))']v^h(t) \right\|^2 \, \mathrm{d}t &\le \int_0^T \|\mathrm{D}^2 F(y_*^h(t))(y_*^h(t))'\|_{\mathrm{op}}^2 \|v^h(t)\|^2 \, \mathrm{d}t \\
&\le \left(\max_{t \in [0,T]} \|v^h(t)\|^2 \right) \int_0^T \|\mathrm{D}^2 F(y_*^h(t))(y_*^h(t))'\|_{\mathrm{op}}^2 \, \mathrm{d}t,
\end{aligned} \tag{39}$$

where $\|\cdot\|_{\mathrm{op}}$ is the operator norm of a matrix (cf. Table 1). To proceed further, note that for any $z \in \mathbb{R}^d$, the Hessian $\mathrm{D}^2 F(z)$ is a bounded linear map from $\mathbb{R}^d$ into $L(\mathbb{R}^d, \mathbb{R}^d)$. We denote by $\left\|\left|\mathrm{D}^2 F(z)\right|\right\|$ the operator norm of $\mathrm{D}^2 F(z)$. Namely,

$$\left\|\left|\mathrm{D}^2 F(z)\right|\right\| = \sup_{w \in \mathbb{R}^d, \|w\|=1} \|\mathrm{D}^2 F(z)w\|_{\mathrm{op}}. \tag{40}$$

Since F is assumed to be C^3, $\left\|\left|\mathrm{D}^2 F(z)\right|\right\|$ is bounded for all z on any bounded set of $\mathbb{R}^d$. Let $\mathfrak{B}$ be the subset in $\mathbb{R}^d$ defined by (14). We get then

$$\begin{aligned}
\int_0^T \|\mathrm{D}^2 F(y_*^h(t))(y_*^h(t))'\|_{\mathrm{op}}^2 \, \mathrm{d}t &\le \sup_{z \in \mathfrak{B}} \left\|\left|\mathrm{D}^2 F(z)\right|\right\|^2 \int_0^T \|(y_*^h(t))'\|^2 \, \mathrm{d}t \\
&\le \sup_{z \in \mathfrak{B}} \left\|\left|\mathrm{D}^2 F(z)\right|\right\|^2 \|y_*^h\|_X^2 \\
&\le \sup_{z \in \mathfrak{B}} \left\|\left|\mathrm{D}^2 F(z)\right|\right\|^2 (r + \|y_*\|_X)^2,
\end{aligned} \tag{41}$$

where the last inequality follows since $y_*^h \in B(y_*, r)$ for all $h \in (0, \bar{h})$; cf. Assumption **(A2)**. Using (41) in (39) and noticing that $\max_{t \in [0,T]} \|v^h(t)\| \le \mathfrak{C} \|v^h\|_X$ (cf. (9)), we get

$$\int_0^T \left\| [\mathrm{D}^2 F(y_*^h(t))(y_*^h(t))']v^h(t) \right\|^2 \, \mathrm{d}t \le \left(\mathfrak{C} \sup_{z \in \mathfrak{B}} \left\|\left|\mathrm{D}^2 F(z)\right|\right\| (r + \|y_*\|_X) \|v^h\|_X \right)^2. \tag{42}$$

Note also that

$$\int_0^T \left\| DF(y_*^h(t))(v^h(t))' \right\|^2 \, dt \le \int_0^T \| DF(y_*^h(t)) \|_{op}^2 \|(v^h(t))'\|^2 \, dt \tag{43}$$
$$\le \sup_{z \in \mathfrak{B}} \| DF(z) \|_{op}^2 \| v^h \|_X^2.$$

By using (42) and (43), we get from (38) that

$$\| \Gamma''(\,\cdot\,; v^h) \|_{L^2(0,T;\mathbb{R}^d)} \le C \| v^h \|_X, \tag{44}$$

where the constant $C > 0$ depends on $\sup_{z \in \mathfrak{B}} \| DF(z) \|_{op}$, $\sup_{z \in \mathfrak{B}} \left\| D^2 F(z) \right\|$, $\| y_* \|_X$ and the embedding constant $\mathfrak{C}$, but is independent of h. The desired estimate (35) follows now from (36), (37) and (44). □

The term $\langle \gamma, v^h - \Pi_h \Gamma(\,\cdot\,; v^h) \rangle_X$ on the LHS of Eq. (34) can be handled using the following lemma.

Lemma 3.5. *Let $\widetilde{\mathfrak{C}} > 0$ be the embedding constant between $L^2(0, T; \mathbb{R}^d)$ and X. If*

$$\widetilde{\mathfrak{C}} \sup_{z \in \mathfrak{B}} \| DF(z) \|_{op} < 1, \tag{45}$$

then there exists a unique $\widehat{v}^h \in X^h$ satisfying

$$\widehat{v}^h - \Pi_h \Gamma(\,\cdot\,; \widehat{v}^h) = y_*^h - w_*^h[y_*^h]. \tag{46}$$

Moreover, it holds that

$$\| \widehat{v}^h \|_X \le \frac{1}{1 - \widetilde{\mathfrak{C}} \sup_{z \in \mathfrak{B}} \| DF(z) \|_{op}} \| y_*^h - w_*^h[y_*^h] \|_X. \tag{47}$$

Proof. Note that the map $\Psi : v^h \to \Pi_h \Gamma(\,\cdot\,; v^h)$ is a bounded linear map from X^h onto itself. To guarantee the existence of a unique $\widehat{v}^h$ that satisfies (46), we only need to show that $\mathrm{Id}_{X^h} - \Psi$ is invertible. By the definition of Γ in (29), we have

$$\| \Psi(v_h) \|_X = \| \Pi_h \Gamma(\,\cdot\,; v^h) \|_X \le \max_{t \in [0,T]} \| DF(y_*^h(t)) \|_{op} \| v^h \|_{L^2(0,T;\mathbb{R}^d)}$$
$$\le \widetilde{\mathfrak{C}} \sup_{z \in \mathfrak{B}} \| DF(z) \|_{op} \| v^h \|_X, \quad \forall \, v^h \in X^h.$$

We get then

$$\| v_h - \Psi(v_h) \|_X \ge (1 - \widetilde{\mathfrak{C}} \sup_{z \in \mathfrak{B}} \| DF(z) \|_{op}) \| v_h \|_X, \quad \forall \, v^h \in X^h. \tag{48}$$

Since $\widetilde{\mathfrak{C}} \sup_{z \in \mathfrak{B}} \| DF(z) \|_{op} < 1$ by our assumption, it follows that the operator norm of $\mathrm{Id}_{X^h} - \Psi$ is bounded below by $1 - \widetilde{\mathfrak{C}} \sup_{z \in \mathfrak{B}} \| DF(z) \|_{op}$. It is thus indeed invertible. The estimate (47) follows directly from (46) and (48). □

We are now in position to prove Proposition 3.2.

Proof of Proposition 3.2. With $\widehat{v}^h$ chosen so that (46) holds and recall the definition of γ given by (33), we get

$$\langle \gamma, \widehat{v}^h - \Pi_h \Gamma(\,\cdot\,; \widehat{v}^h) \rangle_X = \langle y_*^h - w_*[y_*^h], y_*^h - w_*^h[y_*^h] \rangle_X. \tag{49}$$

By rewriting $y_*^h - w_*[y_*^h]$ as $y_*^h - w_*[y_*^h] = \left(y_*^h - w_*^h[y_*^h] \right) + \left(w_*^h[y_*^h] - w_*[y_*^h] \right)$, and recalling from (24) that $w_*^h[y_*^h] - w_*[y_*^h] = -\Pi_h^\perp w_*[y_*^h]$ lives in the orthogonal complement of X^h, we obtain

$$\langle y_*^h - w_*[y_*^h], y_*^h - w_*^h[y_*^h] \rangle_X = \langle y_*^h - w_*^h[y_*^h], y_*^h - w_*^h[y_*^h] \rangle_X = \| y_*^h - w_*^h[y_*^h] \|_X^2.$$

Using this identity in (49), we get

$$\langle \gamma, \widehat{v}^h - \Pi_h \Gamma(\,\cdot\,; \widehat{v}^h) \rangle_X = \| y_*^h - w_*^h[y_*^h] \|_X^2. \tag{50}$$

Note also that by taking $v^h = \widehat{v}^h$ in (34), we have

$$\langle \gamma, \widehat{v}^h - \Pi_h \Gamma(\,\cdot\,; \widehat{v}^h) \rangle_X = \langle \Pi_h^\perp \gamma, \Pi_h^\perp \Gamma(\,\cdot\,; \widehat{v}^h) \rangle_X. \tag{51}$$

Now, it follows from (50), (51), and (35) that

$$\| y_*^h - w_*^h[y_*^h] \|_X^2 \le C \| (w_*[y_*^h])'' \|_{L^2(0,T;\mathbb{R}^d)} \| \widehat{v}^h \|_X h^2. \tag{52}$$

Recall also from Lemma 3.5 that

$$\| \widehat{v}^h \|_X \le \frac{1}{1 - \widetilde{\mathfrak{C}} \sup_{z \in \mathfrak{B}} \| DF(z) \|_{op}} \| y_*^h - w_*^h[y_*^h] \|_X. \tag{53}$$

The desired estimate (26) follows from (52) and (53). □

In the proof of the main theorem given below, we require an upper bound of the term $\|(w_*[y_*^h])''\|_{L^2(0,T;\mathbb{R}^d)}$ appearing on the RHS of (26). This bound should be furthermore independent of the lsfem solution y_*^h. We derive now such a bound. For this purpose, we make use of the solution to the IVP

$$
\begin{aligned}
w' &= F(y_*(t)) + f(t), \quad t \in (0, T), \\
w(0) &= g.
\end{aligned}
\tag{54}
$$

Lemma 3.6. *Let r, $\overline{h}$, and $\mathfrak{B}$ be as specified in Assumption* **(A2)**. *Let $w_*[y_*]$ be the solution of* (54). *Then, for each $h \in (0, \overline{h})$, the solution $w_*[y_*^h]$ of the problem* (23) *can be estimated as:*

$$
\|(w_*[y_*^h])''\|_{L^2(0,T;\mathbb{R}^d)} \le \|(w_*[y_*])''\|_{L^2(0,T;\mathbb{R}^d)} + \sup_{z \in \mathfrak{B}} \|DF(z)\|_{\mathrm{op}}\Big(r + 2\|y_*\|_X\Big).
\tag{55}
$$

Proof. Note that by using the triangle inequality,

$$
\|(w_*[y_*^h])''\|_{L^2(0,T;\mathbb{R}^d)} \le \|(w_*[y_*^h] - w_*[y_*])''\|_{L^2(0,T;\mathbb{R}^d)} + \|(w_*[y_*])''\|_{L^2(0,T;\mathbb{R}^d)},
\tag{56}
$$

we only need to estimate the term $\|(w_*[y_*^h] - w_*[y_*])''\|_{L^2(0,T;\mathbb{R}^d)}$. Since $w_*[y_*^h]$ and $w_*[y_*]$ are respectively the solutions to the IVPs (23) and (54), we have

$$
(w_*[y_*^h])' - (w_*[y_*])' = F(y_*^h) - F(y_*).
\tag{57}
$$

Then,

$$
(w_*[y_*^h] - w_*[y_*])'' = DF(y_*^h)(y_*^h)' - DF(y_*)(y_*)',
\tag{58}
$$

which leads to

$$
\begin{aligned}
\| (w_*[y_*^h] &- w_*[y_*])'' \|_{L^2(0,T;\mathbb{R}^d)} \\
&= \| DF(y_*^h)(y_*^h)' - DF(y_*)(y_*)' \|_{L^2(0,T;\mathbb{R}^d)} \\
&\le \|DF(y_*^h)(y_*^h - y_*)'\|_{L^2(0,T;\mathbb{R}^d)} + \|(DF(y_*^h) - DF(y_*))(y_*)'\|_{L^2(0,T;\mathbb{R}^d)} \\
&\le \sup_{z \in \mathfrak{B}} \|DF(z)\|_{\mathrm{op}}\Big(\|(y_*^h - y_*)'\|_{L^2(0,T;\mathbb{R}^d)} + 2\|(y_*)'\|_{L^2(0,T;\mathbb{R}^d)}\Big) \\
&\le \sup_{z \in \mathfrak{B}} \|DF(z)\|_{\mathrm{op}}\Big(\|y_*^h - y_*\|_X + 2\|y_*\|_X\Big) \\
&\le \sup_{z \in \mathfrak{B}} \|DF(z)\|_{\mathrm{op}}\Big(r + 2\|y_*\|_X\Big).
\end{aligned}
\tag{59}
$$

In deriving (59), we have used the facts that $\|y_*^h - y_*\|_X < r$ since $y_*^h \in B(y_*, r)$ and that $\mathfrak{B}$ contains both $y_*(t)$ and $y_*^h(t)$ for all $t \in [0, T]$; see (14) and (15). The desired result (55) follows from (59) and (56). $\quad\square$

We are now in position to prove the main theorem of this section.

Proof of Theorem 3.1. First note that by the triangle inequality, we get

$$
\|y_* - y_*^h\|_{L^2(0,T;\mathbb{R}^d)} \le \|y_* - w_*[y_*^h]\|_{L^2(0,T;\mathbb{R}^d)} + \|w_*[y_*^h] - w_*^h[y_*^h]\|_{L^2(0,T;\mathbb{R}^d)} + \|w_*^h[y_*^h] - y_*^h\|_{L^2(0,T;\mathbb{R}^d)}.
\tag{60}
$$

To estimate the first term $\|y_* - w_*[y_*^h]\|_{L^2(0,T;\mathbb{R}^d)}$ on the RHS of (60), we integrate Eqs. (3) and (23) to obtain

$$
\begin{aligned}
\|y_*(t) - w_*[y_*^h](t)\| &\le \int_0^t \|F(y_*(s)) - F(y_*^h(s))\| \, \mathrm{d}s \\
&\le \mathrm{Lip}(F|_{\mathfrak{B}}) \int_0^t \|y_*(s) - y_*^h(s)\| \, \mathrm{d}s \\
&\le \mathrm{Lip}(F|_{\mathfrak{B}})\sqrt{t} \left(\int_0^t \|y_*(s) - y_*^h(s)\|^2 \, \mathrm{d}s \right)^{1/2}, \quad t \in [0, T],
\end{aligned}
\tag{61}
$$

where we applied Hölder's inequality in the last step above. We get in turn that

$$
\|y_* - w_*[y_*^h]\|_{L^2(0,T;\mathbb{R}^d)} \le \frac{\mathrm{Lip}(F|_{\mathfrak{B}})T}{\sqrt{2}} \|y_* - y_*^h\|_{L^2(0,T;\mathbb{R}^d)}.
\tag{62}
$$

The second term $\|w_*[y_*^h] - w_*^h[y_*^h]\|_{L^2(0,T;\mathbb{R}^d)}$ in (60) can be estimated by using (25), and the last term $\|w_*^h[y_*^h] - y_*^h\|_{L^2(0,T;\mathbb{R}^d)}$ can be estimated by using (26) together with

$$
\|w_*^h[y_*^h] - y_*^h\|_{L^2(0,T;\mathbb{R}^d)} \le \widetilde{\mathfrak{C}}\|w_*^h[y_*^h] - y_*^h\|_X,
\tag{63}
$$

where $\widetilde{\mathfrak{C}}$ denotes again the embedding constant between $L^2(0, T; \mathbb{R}^d)$ and X.

Gathering the above estimates for the three terms on the RHS of (60), we get

$$\|y_* - y_*^h\|_{L^2(0,T;\mathbb{R}^d)} \leq \frac{\text{Lip}(F|_{\mathfrak{B}})T}{\sqrt{2}} \|y_* - y_*^h\|_{L^2(0,T;\mathbb{R}^d)} + C\left(1 + \frac{1}{1 - \widetilde{\mathfrak{C}} \sup_{z \in \mathfrak{B}} \|DF(z)\|_{\text{op}}}\right) \|(w_*[y_*^h])''\|_{L^2(0,T;\mathbb{R}^d)} h^2, \qquad (64)$$

where we have absorbed the factor $\widetilde{\mathfrak{C}}$ on the RHS of (63) into the constant C when applying the estimate (26).

In the above inequality, by using the estimate (55) for $\|(w_*[y_*^h])''\|_{L^2(0,T;\mathbb{R}^d)}$, we get after rearranging terms that

$$\left(1 - \frac{\text{Lip}(F|_{\mathfrak{B}})T}{\sqrt{2}}\right) \|y_* - y_*^h\|_{L^2(0,T;\mathbb{R}^d)} \leq \widetilde{C} h^2, \qquad (65)$$

where

$$\widetilde{C} := C\left(1 + \frac{1}{1 - \widetilde{\mathfrak{C}} \sup_{z \in \mathfrak{B}} \|DF(z)\|_{\text{op}}}\right) \left(\|(w_*[y_*])''\|_{L^2(0,T;\mathbb{R}^d)} + \sup_{z \in \mathfrak{B}} \|DF(z)\|_{\text{op}}(r + 2\|y_*\|_X)\right), \qquad (66)$$

with $w_*[y_*]$ denoting the solution of (54). Since it is assumed that $\frac{\text{Lip}(F|_{\mathfrak{B}})T}{\sqrt{2}} < 1$ (see (16)), the desired result (18) follows by taking C therein to be $\widetilde{C}/(1 - \text{Lip}(F|_{\mathfrak{B}})T/\sqrt{2})$. $\quad\square$

4. Numerics

In this section we focus on numerical aspects of lsfem by discussing algorithmic details and confirming the analytical insight gained in previous Sections 2 and 3 through numerical experiments. We discuss modifications such as adaptive time stepping and constrained systems, i.e., differential algebraic equations. Our proposed method is summarized in Algorithm 1.

Algorithm 1 (lsfem for IVPs)

1: **function** y_*^h = lsfem($G(\cdot, \cdot)$, g, $[t_0, T]$, X^h)
2:     construct finite element basis for $y^h(\cdot; \cdot)$ and $(y^h)'(\cdot; \cdot)$ of X^h
3:     choose discretization $\underline{t}$ of time interval
4:     compute

$$x_* \in \underset{x}{\text{argmin}} \ \mathcal{J}(x) = \tfrac{1}{2}\|(y^h)'(\underline{t}; x) - G(\underline{t}, y^h(\underline{t}; x))\|^2 + \tfrac{1}{2}\|y^h(0; x) - g\|^2$$

5:   set $y_*^h = y^h(\cdot; x_*)$
6: **end function**

A Matlab implementation of Algorithm 1 with equidistant control points is available at github.com/matthiaschung/lsfem and is intended for developing an understanding of the performance of lsfem method for ODEs. As inputs lsfem requires the RHS of the first order ODE G, the interval of interest $[t_0, T]$, and the initial condition $y(t_0) = g$,

$$\begin{aligned} y' &= G(t, y), \quad t \in (t_0, T], \\ y(t_0) &= g. \end{aligned} \qquad (67)$$

Additionally, one may select a desired finite element space X^h. Algorithm 1 return the function $y_h^* = y_h(\cdot; x_*)$ determined by the optimized finite element coefficients x_* with respect to the corresponding finite element basis.

Various numerical choices need to be made in Algorithm 1. First, the approximation quality of our method depends greatly on the choice of the finite element space X^h and its corresponding control points τ in line 2. Sections 2 and 3 provides convergence results for piecewise linear basis functions, however, we may choose higher order basis functions. Common choices for the basis function include piecewise polynomials and polynomial splines [20,21]. Other options are exponential splines, which may better capture the exponential behavior manifested by certain differential equations; see [22]. An interesting alternative are Hermite splines [23], which are able to take advantage of derivative information provided naturally by the differential equation and reducing computational costs. Note that, the choice of the finite element basis may depend on the imposed smoothness of the underlying dynamical system, i.e., G. Equidistant control points may be selected if no further information on y are available, however, one may also select control points if knowledge on y (or its derivatives) are available.

To numerically evaluate and minimize J of Eq. (6), quadrature is required to approximate the L^2-norm. Hence, with respect to the quadrature rule we discretize the interval $[t_0, T]$ with $\underline{t} = [t_0, t_1, \ldots, t_{n-1}, T]^\top$, see line 3. Choosing a quadrature rule (such as Gauss–Legendre and Gauss–Lobatto, [24]) which is consistent with the finite element space X^h may provide computational advantages. The added benefit of using such a quadrature rule is that the resulting ℓ^2-norm approximation $\mathcal{J}$ has the potential to be exact in certain polynomial settings.
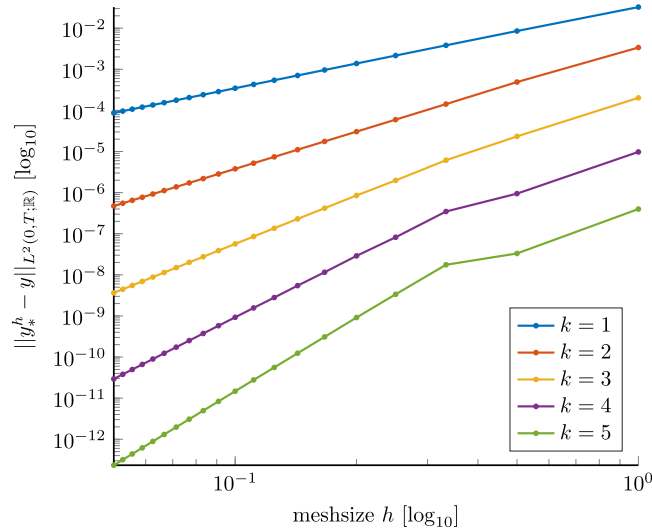
Fig. 2. Illustration of convergence rates of `lsfem` methods for B-spline bases of degree k ($k = 1, \ldots, 5$) in comparison to the mesh sizes h for the ODE $y' = -y$ with $t \in [0, 1]$ and $y(0) = 1$. The numerically observed convergence rates are 1.9972 (for $k = 1$), 3.0066 (for $k = 2$), 3.9204 (for $k = 3$), 4.9456 (for $k = 4$) and 5.9409 (for $k = 5$).

The main computational effort lies in line 4. Line 4 defines a common (regularized) nonlinear least squares problem. Notice that, if G is sufficiently smooth, gradient and Hessian based optimization methods can be utilized.

Hence, gradient based methods and also Newton type methods are natural choices (assuming sufficient smoothness of the system). However, $\nabla_x J$ and $\nabla_x^2 J$ need to be readily available or be obtained by algorithmic differentiation techniques [25]. It is worth mentioning, that `lsfem` seeks for a global minimizer y_h^* of (6). However, for non-convex problems the proposed optimization methods may not ensure convergence to the global minimizer. Strategies to prevent local minimizer are required, e.g., multi-start or global optimization methods [26].

Rate of convergence for higher order finite elements. To illustrate and empirically valid the convergence rates discussed in Sections 2 and 3, we first consider the linear initial value problem $y' = -y$, with $t \in [0, 1]$ and $y(0) = 1$. We use a B-spline bases for X^h of degree $k = 1, \ldots, 5$ with varying equidistant discretization of the finite elements, i.e., $h = 1/N$ with $N = 1, \ldots, 20$ to compute the finite element approximation y_*^h. Fig. 2 depicts the errors $\|y_*^h - y\|$ with respect to the varying mesh sizes h in a log–log space. The slopes of each graph reveal the power of the expected convergence rates of our method. For instance the slope using linear B-splines is 1.9972 confirming the quadratic convergence rate (Theorem 3.1). The other rates are 3.0066 (for $k = 2$), 3.9204 (for $k = 3$), 4.9456 (for $k = 4$) and 5.9409 (for $k = 5$), respectively. These results lead us to conjecture that the optimal `lsfem` error bounds scale like $\mathcal{O}(h^{k+1})$ for finite element bases of degree k. By inspecting the proofs in Section 3, we expect many of the ingredients presented there to be extended naturally, although some aspects such as higher-order analogues of Lemma 3.6 may require additional efforts. We plan to address such an extension in a future work.

To confirm the results for nonlinear ODEs we consider the simple logistic growth $y' = y(1 - y)$ with $y(0) = 1/10$ and $t \in [0, 10]$. We compare our `lsfem` method with B-spline bases of degree 1, 2, and 3 (`lsfem1`, `lsfem2`, and `lsfem3`) to Runge–Kutta 3 (`rk3`) and Runge–Kutta 4 (`rk4`), see Fig. 3. The numerically observed convergence rates for this logistic growth model are 2.001, 3.4928, and 4.0470 for the `lsfem` methods and 2.9731, 3.9820 for the Runge–Kutta methods, respectively. The observed rates for `lsfem` confirm again the obtained theoretical estimate for degree $k = 1$ case and corroborate the conjectured optimal bound $\mathcal{O}(h^{k+1})$ for higher-degree bases (with `lsfem2` providing actually better rate than conjectured for this particular example). One can also compare `lsfem3` with `rk4`, since both methods show a convergence rate close to the theoretical rate $r = 4$. Fig. 3 reveals that the constant C in the associated error bound Ch^4 is smaller in the case of `lsfem3` than that of `rk4` for the considered example.

Linear ODEs. In case of linear ODEs the `lsfem`'s main computational burden of solving the optimization problem in line 4 of Algorithm 1 simplifies to a linear least-squares problem whose solution can be obtained e.g., by solving the associated linear normal equations.

More precisely, let us consider the n dimensional initial value problem

$$y'(t) = A(t)y(t) + b(t) \quad \text{and} \quad y(t_0) = g \text{ and } t \in (t_0, T]. \tag{68}$$

Assuming we choose the same finite element basis for each state

$$\phi(t) = [\phi_1(t), \phi_2(t), \ldots, \phi_m(t)]^\top, \tag{69}$$
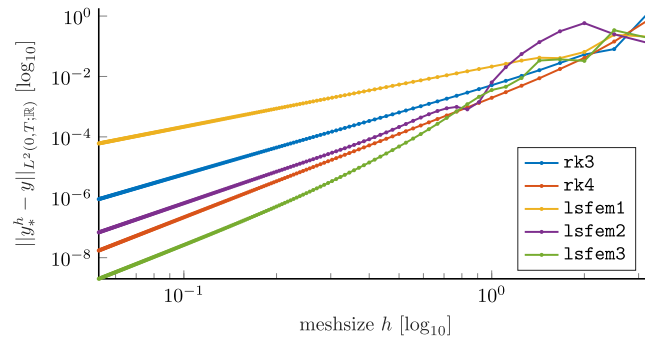
Fig. 3. Illustration of convergence rates of `lsfem` utilizing B-spline bases of degree 1, 2, and 3 in comparison to Runge–Kutta 3 and 4 for the logistic equation $y' = y(1 - y)$, with $y(0) = 1/10$ and $t \in [0, 10]$. The numerically observed convergence rates are 2.001, 3.4928, and 4.0470 for the `lsfem` methods and 2.9731, 3.9820 for the Runge–Kutta methods, respectively.
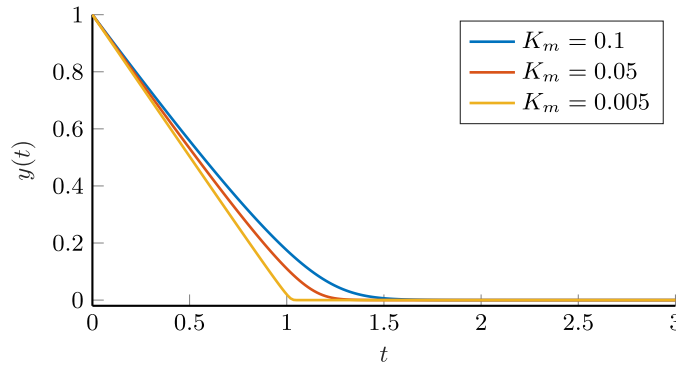
Fig. 4. Solution of Eq. (74) for three different values of K_m and with $y(0) = 1$. By decreasing K_m the rapid change of dynamics of y become more prominent around $t = 1$.

then the function of the finite element space are given by $y^h(t, x) = (\phi(t)^\top \otimes I_n)x$ with some coefficients

$$x = \left[x_1^1, \ldots, x_1^n, x_2^1, \ldots, x_2^n, \ldots, x_m^1, \ldots, x_m^n\right]^\top \tag{70}$$

where $\otimes$ denotes the Kronecker product and I_n the identity matrix. The least-squares problem now reads

$$\min_x \ \frac{1}{2} \int_{t_0}^T r(t, x)^\top r(t, x) \, dt + \frac{1}{2} \left\| y^h(t_0, x) - g \right\|_2^2 \tag{71}$$

with $r(t, x) = Z(t)x - b(t)$, where $Z(t) = \phi'(t)^\top \otimes I_n - A(t)(\phi(t)^\top \otimes I_n)$. With the further abbreviations

$$Q = \int_{t_0}^T Z(t)^\top Z(t) \, dt, \quad p = \int_{t_0}^T Z(t)^\top b(t) \, dt, \quad \text{and} \quad R = \phi(t_0)^\top \otimes I_n. \tag{72}$$

The `lsfem` solution of (68) is obtained by the normal equations

$$(Q + R^\top R)x_* = p + R^\top g, \tag{73}$$

and $y_*^h(t) = (\phi(t)^\top \otimes I_n)(Q + R^\top R)^{-1}(p + R^\top g)$, assuming $Q + R^\top R$ is invertible. Hence standard linear algebra libraries may be utilized to solve a linear system of differential equations efficiently.

Adaptive discretization of the finite elements. So far we have not discussed how to select control points τ of our finite element space X^h and assumed they are pre-selected, e.g., equidistant. Alternatively, control points may be selected adaptively by (for simplicity) repeated evaluation of line 4 with refined control points τ. Notice, `lsfem` naturally provides error estimates through the residuals $r = (y^h)'(\underline{t}; x_*) - G(\underline{t}, y^h(\underline{t}; x_*))$. In its simplest form new control points τ_i's may be introduced by selecting discretization points $\underline{t}_j$'s at locations with large residuals r_j.

We illustrate this procedure by considering the model

$$y' = -\frac{y}{K_m + y}, \tag{74}$$
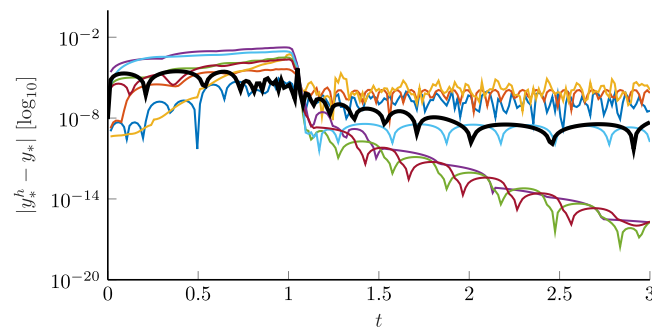
Fig. 5. Absolute error between various ODE methods and the true solution for Eq. (74) with $K_m = 0.005$ and $y(0) = 1$. Our `lsfem` method is highlighted in black bold, other standard Matlab ODE solvers in default settings are illustrated as labeled in Fig. 6.
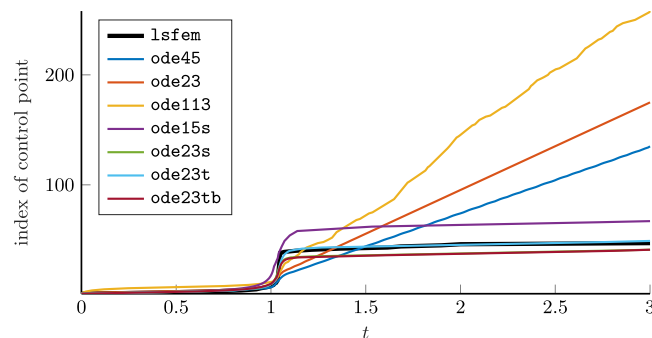
Fig. 6. Time location of the corresponding control point for `lsfem` (black bold) in comparison with time location of the discretization index of FDMs, for Eq. (74) with $K_m = 0.005$ and $y(0) = 1$. Notice that all methods, especially `lsfem` and all the stiff FDM solvers, select small step sizes around $t = 1$ where the kink in the true solution occurs (cf. Fig. 4). At the same time, the three non-stiff solvers `ode45`, `ode23`, and `ode113` also require a fine discretization beyond $t = 1$ even though the solution dynamics is "quiescent" there (see again Fig. 4). Note also that the curve for `ode23s` almost overlaps with `ode23tb`. All Matlab ODE solvers are in default settings with default error controls.

with $y(0) = 1$, $t \in [0, 3]$. The solution is implicitly given by $y + K_m \ln y = 1 - t$. Eq. (74) is a simplified model frequently appearing in the field of enzyme kinetics, [27,28]. Here $K_m > 0$ refers to the Michaelis–Menten constant determining the reaction rate. For small K_m, e.g., $K_m = 0.005$, stiff ODE (74) exhibits a decay with a sharp "kink" around $t = 1$, see Fig. 4, resulting in difficulties for numerical ODE solvers to preserve non-negative concentrations. Adaptive refinement of this particular area is crucial. We initialize `lsfem` with four equidistant control points (order 3 and 8 Gauss Legendre points) and refine the control points τ of our finite element basis until each residual element reaches an absolute tolerance of 10^{-4}. The error of `lsfem` with respect to the true solution and in comparison of standard Matlab is depicted in Fig. 5 while Fig. 6 shows the number of discretization points vs. the location of these points. We observe that `lsfem` adaptively adds control points around 1 and maintains an absolute error below $5 \cdot 10^{-5}$ throughout the time interval while requiring 47 control points.

In comparison, stiff ODE solvers such as `ode15s`, `ode23s`, `ode23t`, and `ode23tb` also adaptively refine around $t = 1$ and the total time steps used are similar to the number of control points for `lsfem` except `ode15s`, with the latter taking more time steps than the other stiff solvers as shown in Fig. 6. Note also that the curve for `ode23s` almost overlaps with `ode23tb` in Fig. 6 and is thus not visible. In terms of errors, the `lsfem` performs better over the interval $[0, 1]$ in which the dynamics is "non-trival", and the stiff solvers perform better over the interval $[1, 3]$ in which the dynamics is "quiescent"; see Figs. 4 and 5. Overall, `lsfem` performs similar to stiff FDM solvers, while maintaining a slightly lower L^2-error. Meanwhile, non-stiff ODE solvers such as `ode23`, `ode45`, and `ode113` refine less around $t = 1$ but need a significantly more number of discretization points beyond $t > 1$ to maintain numerical accuracy.

Finally, as another example to demonstrate the role of adaptive discretization, we consider the unforced Van der Pol oscillator, which evolves according to the second-order differential equation

$$y'' - c(1 - y^2)y' + y = 0, \tag{75}$$

where $c > 0$ is a parameter. It is known that (75) admits a unique stable limit cycle for any $c > 0$ [29, Section 3.8]. As c increases, the equation becomes more and more stiff due to the presence of sharp gradients in the limit cycle; see panels (a)–(c) of Fig. 7.

To apply `lsfem` (and for comparison `ode15s`), we rewrite (75) as a first-order system for (y, z) where $z = y'$. In panels (a)–(c) of Fig. 7, we show by red curves the solution y obtained from `lsfem` for three different c values, with the initial
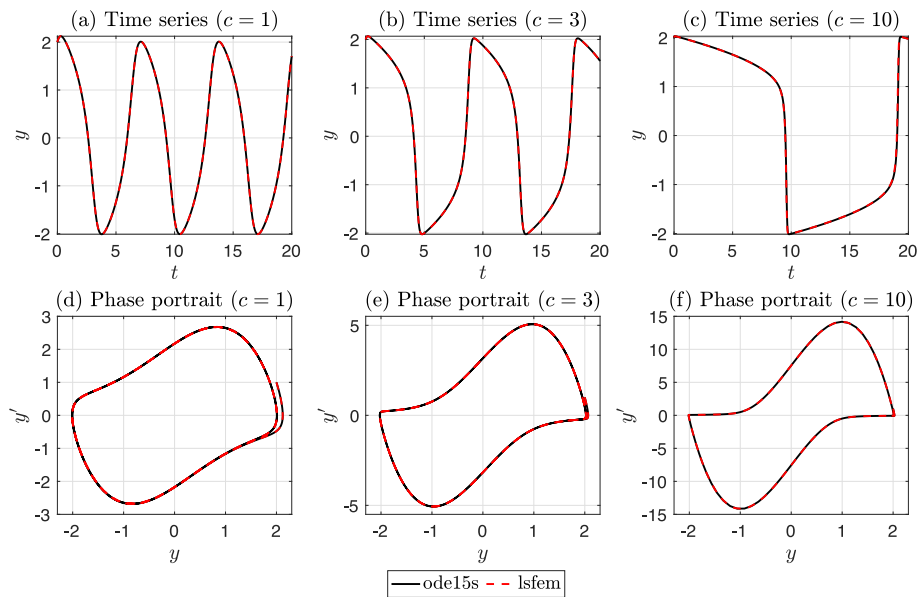
Fig. 7. Panels (a)–(c): The `lsfem` solution of the Van der Pol oscillator (75) for three different values of the parameter c and with initial data $(y(0), y'(0)) = (2, 1)$ (red dashed curve). Panels (d)–(f): The phase portraits in the coordinates (y, y') for the corresponding solutions shown in panels (a)–(c). A high-precision reference solution obtained from `ode15s` is also shown for each case (black solid curve).

data fixed to be $(y(0), y'(0)) = (2, 1)$. The corresponding phase portraits in the coordinates (y, y') are shown in panels (d)–(f) of Fig. 7. Like the enzyme kinetics model above, the adaptive refinement in `lsfem` for the Van der Pol oscillator automatically allocates more control points in the time windows where sharp gradients take place, since sharper gradients typically lead to larger residual values when the control points located around the sharp gradient are not dense enough. For the results presented in Fig. 7, the `lsfem`'s adaptive refinement is terminated when the residual at all control points are below 10^{-4} for both y and y'. Since there do not exist analytic solutions of (75) for general initial data, as a reference solution, we also showed by the black curves the corresponding solutions obtained from the Matlab built-in stiff FDM solver `ode15s`, for which the relative and absolute tolerances are both set to be 10^{-10}. As can be observed, the `lsfem` solution matches well with this high-precision reference solution.

5. Conclusion and discussion

In this work, we considered the least-squares finite element method (`lsfem`) for systems of nonlinear ordinary differential equations and established under suitable conditions an optimal error estimate for this method when piecewise linear elements are used (Theorem 3.1). In contrast to the "localization" nature of finite difference methods, the `lsfem` aims to find an optimal approximate solution within a given subspace that minimizes an objective function over the whole time interval of integration. The `lsfem` can thus be less prone to the accumulation of local discretization errors compared to finite difference methods.

As reviewed in Section 1, a key ingredient in our derivation of the optimal estimate is a geometric (orthogonality) property derived from the first-order optimality condition associated with the minimizers of the underlying optimization problems; see Eq. (2). Numerical results presented in Section 4 not only support our main theoretical result presented in Theorem 3.1, but also provide strong indication that error bound of the form $\mathcal{O}(h^{k+1})$ will hold if higher-order spline basis elements of degree k ($k \geq 2$) were used. In Section 4, we also discussed details related to the associated algorithmic aspects (Algorithm 1) as well as suitable modifications for adaptive mesh refinement to handle ODEs whose solutions may experience abrupt local changes. It is also worth mentioning that residual neural networks appear to be of such nature and we will dedicate future research towards such applications, [9,10].

Finally, we mention that the procedure presented in Algorithm 1 can be easily adapted to handle a broad class of differential algebraic equations (DAEs) [4,30,31] as well. One just needs to add the corresponding algebraic equations as constraints to the associated optimization problems. The numerical setup can also be easily extended to handle ODE boundary value problems and delay differential equations. On the application side, due to its robustness in handling stiff problems, we anticipate the `lsfem` to be effective for e.g., low dimensional *parameterized* systems, such as those frequently appearing in parameter estimation problems for systems biology applications [32–35], in which the underlying ODE system can be either stiff or non-stiff as the parameters are varied. Those extensions and applications are beyond the scope of this work and will remain for future communications.

Appendix A. Proof of Theorem 2.1

We start by rewriting the respective first-order optimality condition associated with the minimization problems (5) and (6) into an abstract form $u + \mathcal{T} \circ \mathcal{G}(u) = 0$ for (5) and $u^h + \Pi_h \mathcal{T} \circ \mathcal{G}(u^h) = 0$ for (6), where $\mathcal{T}$ is a bounded linear operator and $\mathcal{G}$ is a smooth nonlinear operator defined below, and Π_h is the orthogonal projection onto X^h appearing in (11). These equations are in the same functional forms dealt with in [18, Theorem 3.3, p.307]. Once this reformulation is done, we just need to check that all conditions required in [18, Theorem 3.3, p.307] are satisfied and that the solutions for $u^h + \Pi_h \mathcal{T} \circ \mathcal{G}(u^h) = 0$ guaranteed by [18, Theorem 3.3, p.307] are indeed minimizers of (6). The reformulation is dealt with in Lemma Appendix A.1, and the remaining steps are carried out afterwards.

As a preparation, we define the linear operator $\mathcal{T}$ to be the solution operator associated with the special case of (3) in which F is identically zero. That is,

$$\mathcal{T} : L^2(0, T; \mathbb{R}^d) \times \mathbb{R}^d \to X, \ \mathcal{T}(f, g) = \ell, \ \text{with } \ell(t) = g + \int_0^t f(s)\, \mathrm{d}s, \ \forall\, t \in [0, T]. \tag{A.1}$$

Note that $\mathcal{T}$ thus defined is obviously a bounded linear operator. It follows directly from the Riesz representation theorem (see e.g., [36, Appendix D] and [37, Section 16.2]), that $\ell = \mathcal{T}(f, g)$ defined above is the unique element in X satisfying

$$\langle \ell, v \rangle_X = \int_0^T \langle f(t), v'(t) \rangle\, \mathrm{d}t + \langle g, v(0) \rangle, \qquad \forall\, v \in X. \tag{A.2}$$

Note also that this identity is the first-order necessary condition for ℓ to be a solution of (5) when F is absent; cf. (A.8).

Assume that F is C^1 smooth. For any (f, g) in $L^2(0, T; \mathbb{R}^d) \times \mathbb{R}^d$, we define now a nonlinear operator $\mathcal{G}$ as follows:

$$\mathcal{G} : X \to L^2(0, T; \mathbb{R}^d) \times \mathbb{R}^d, \qquad \mathcal{G}(u) = (\widetilde{f}(u), \widetilde{g}(u)), \tag{A.3}$$

where for each u in X, $(\widetilde{f}(u), \widetilde{g}(u))$ is defined by

$$
\begin{aligned}
& [\widetilde{f}(u)](t) = -F(u(t)) - f(t) - \int_t^T [DF(u(s))]^\top \left(u' - F(u) - f(s) \right) \mathrm{d}s, \ \forall\, t \in [0, T], \\
& \widetilde{g}(u) = -g - \int_0^T [DF(u(s))]^\top \left(u' - F(u) - f(s) \right) \mathrm{d}s.
\end{aligned}
\tag{A.4}
$$

To see that $(\widetilde{f}(u), \widetilde{g}(u))$ thus defined is indeed an element in $L^2(0, T; \mathbb{R}^d) \times \mathbb{R}^d$, note that since $u \in X$ and F is assumed to be C^1, we have $w = u' - F(u) - f \in L^2(0, T; \mathbb{R}^d)$ and $DF(u)$ is continuous on $[0, T]$. One can then show that $\int_0^T (DF(u))^\top w \, \mathrm{d}s$ is finite and the function $t \mapsto \psi(t) := \int_t^T (DF(u))^\top w \, \mathrm{d}s$ is in $L^2(0, T; \mathbb{R}^d)$. As a result, $\widetilde{f}$ maps X into $L^2(0, T; \mathbb{R}^d)$ and $\widetilde{g}$ maps X into $\mathbb{R}^d$.

The rationale behind the definition of $\mathcal{G}$ will become apparent in the proof of the following lemma.

Lemma Appendix A.1. *Consider the IVP (3). Assume that $f \in L^2(0, T; \mathbb{R}^d)$ and $F : \mathbb{R}^d \to \mathbb{R}^d$ is C^1. Let $\mathcal{T}$ and $\mathcal{G}$ be defined by (A.1) and (A.3), respectively. Then, any strong solution of (3) also satisfies the following nonlinear problem*

$$u + \mathcal{T} \circ \mathcal{G}(u) = 0. \tag{A.5}$$

Similarly, denoting by $\Pi_h : X \to X^h$ the orthogonal projection onto X_h, any solution of (6) also satisfies

$$u^h + \Pi_h \mathcal{T} \circ \mathcal{G}(u^h) = 0. \tag{A.6}$$

Proof. We organize the proof into two steps. Step 1 deals with the original problem (3); and Step 2 deals with its `lsfem` formulation.

Step 1. Note that since y_* solves (3), it is a minimizer of the objective function J given by (4). As a result, y_* satisfies the following first-order necessary condition:

$$\left. \frac{\mathrm{d}}{\mathrm{d}\tau} J(y_* + \tau v; F, f, g) \right|_{\tau=0} = 0, \qquad \forall\, v \in X. \tag{A.7}$$

By using (4) in (A.7), we obtain the following integral equation to be satisfied by y_*:

$$\int_0^T \langle u' - F(u) - f, v' - DF(u)v \rangle\, \mathrm{d}t + \langle u(0) - g, v(0) \rangle = 0, \qquad \forall\, v \in X, \tag{A.8}$$

where DF denotes the Jacobian matrix of F.

In order to rewrite (A.8) into the form $u + \mathcal{T} \circ \mathcal{G}(u) = 0$, let us introduce:

$$Q(u, v; f, g) = \int_0^T \langle u' - F(u) - f, DF(u)v \rangle \, dt + \int_0^T \langle F(u) + f, v' \rangle \, dt + \langle g, v(0) \rangle, \tag{A.9}$$

which is defined for any u, v in X, f in $L^2(0, T; \mathbb{R}^d)$ and g in $\mathbb{R}^d$. Note that, with the above definition of Q and the definition of the inner product on X given by (7), (A.8) can be rewritten as

$$\langle u, v \rangle_X - Q(u, v; f, g) = 0, \quad \forall v \in X. \tag{A.10}$$

To proceed further, we note that $Q(u, v; f, g)$ defined in (A.9) satisfies

$$Q(u, v; f, g) + \int_0^T \langle \widetilde{f}(u), v' \rangle \, dt + \langle \widetilde{g}(u), v(0) \rangle = 0, \qquad \forall u, v \in X, \tag{A.11}$$

where $(\widetilde{f}(u), \widetilde{g}(u))$ is an element in $L^2(0, T; \mathbb{R}^d) \times \mathbb{R}^d$ given by (A.4).

The above identity (A.11) can be derived from integrating by parts the first term in the definition of $Q(u, v; f, g)$ given in (A.9). Indeed, denoting $w = u' - F(u) - f$ and by noting that $\langle u' - F(u) - f, DF(u)v \rangle = \langle w, DF(u)v \rangle = (DF(u)v)^\top w = v^\top (DF(u))^\top w$, we have

$$\begin{aligned}
\int_0^T \langle w, DF(u)v \rangle \, dt &= \int_0^T v^\top (DF(u))^\top w \, dt \\
&= -\int_0^T v^\top \, d\left(\int_t^T (DF(u))^\top w \, ds \right) \\
&= \left\langle \int_0^T (DF(u))^\top w \, ds, v(0) \right\rangle + \int_0^T \left\langle \int_t^T (DF(u))^\top w \, ds, v' \right\rangle \, dt,
\end{aligned} \tag{A.12}$$

where we used integration by parts to obtain the last equality above. Now, replacing the first term on the RHS of (A.9) with the RHS of (A.12) and using the definition of $(\widetilde{f}(u), \widetilde{g}(u))$ given by (A.4), we obtain (A.11).

Thanks to the identity (A.11), an element u in X satisfies (A.10) if and only if

$$\langle u, v \rangle_X = \int_0^T \langle -\widetilde{f}(u), v' \rangle \, dt + \langle -\widetilde{g}(u), v(0) \rangle, \quad \forall v \in X. \tag{A.13}$$

Recalling the equivalent characterization given in (A.2) of the linear operator $\mathcal{T}$ defined by (A.1), we get from (A.13) that

$$u = \mathcal{T}(-\widetilde{f}(u), -\widetilde{g}(u)), \text{ or equivalently, } u + \mathcal{T}(\widetilde{f}(u), \widetilde{g}(u)) = 0.$$

At the same time, by the definition of $\mathcal{G}$ in (A.3), we have $(\widetilde{f}(u), \widetilde{g}(u)) = \mathcal{G}(u)$. Using this relation in $u + \mathcal{T}(\widetilde{f}(u), \widetilde{g}(u)) = 0$, we obtain the desired form $u + \mathcal{T} \circ \mathcal{G}(u) = 0$ for the first-order optimality condition (A.8).

Step 2. Now, we consider the `lsfem` problem (6) which aims to approximate the variational formulation (5) of (3). Note that any solution y_*^h of (6), if exists, satisfies the analogue of (A.8) with v therein restricted to X^h. That is

$$\int_0^T \langle (y_*^h)' - F(y_*^h) - f, (v^h)' - DF(y_*^h)v^h \rangle \, dt + \langle y_*^h(0) - g, v^h(0) \rangle = 0, \ \forall v^h \in X^h. \tag{A.14}$$

We can then follow the same derivation of (A.5) from (A.8) to obtain that y_*^h is a solution of the following nonlinear problem defined on X^h:

$$u^h + \mathcal{T}^h \circ \mathcal{G}(u^h) = 0, \quad u^h \in X^h, \tag{A.15}$$

where $\mathcal{G}$ is the same as defined in (A.3), and $\mathcal{T}^h : L^2(0, T; \mathbb{R}^d) \times \mathbb{R}^d \to X^h$ is defined by

$$\mathcal{T}^h(f, g) = u^h \quad \text{if and only if}$$
$$\langle u^h, v^h \rangle_X = \int_0^T \langle f, (v^h)' \rangle \, dt + \langle g, v^h(0) \rangle, \qquad \forall v^h \in X^h. \tag{A.16}$$

Comparing (A.15) with (A.6), it remains to show that

$$\mathcal{T}^h = \Pi_h \mathcal{T}. \tag{A.17}$$

To see this, for any $(f, g) \in L^2(0, T; \mathbb{R}^d) \times \mathbb{R}^d$, we get from (A.2) and (A.16) that

$$\langle \mathcal{T}(f, g) - \mathcal{T}^h(f, g), v \rangle_X = 0, \quad \forall v \in X^h.$$

Namely, $\mathcal{T}(f, g) - \mathcal{T}^h(f, g)$ belongs to the orthogonal complement of X^h; that is $\mathcal{T}^h = \Pi_h \mathcal{T}$. The proof is complete. $\quad \square$

Thanks to Lemma Appendix A.1, we have thus reformulated the first-order optimality condition associated with each of the minimization problems (5) and (6) into the desired form given by (A.5) and (A.6), respectively.

Note that (A.5) and (A.6) fit into the abstract formulation of [18, Theorem 3.3, p.307] (see also [12, Theorem 8.1]). For a given nonsingular solution y_* of (A.5), [18, Theorem 3.3, p.307] delineates conditions to ensure the existence of a solution for (A.6) for all sufficiently small h that converges to y_*.[2]

A solution y_* to (A.5) is called nonsingular if the linear operator $\mathrm{Id}_X + \mathcal{T} \circ \mathrm{D}\mathcal{G}(y_*) : X \to X$ is invertible with bounded inverse; namely,

$$(\mathrm{Id}_X + \mathcal{T} \circ \mathrm{D}\mathcal{G}(y_*))^{-1} \in L(X, X), \tag{A.18}$$

where Id_X denotes the identity map on X, and $L(X, X)$ the set of all bounded linear operators on X. As will be shown below that, for the problem at hand, the condition (A.18) is ensured by the smallness assumption on the operator norm of the Jacobian matrix $\mathrm{D}F(y_*(t))$ for all t in $[0, T]$.

Proof of Theorem 2.1. From what precedes, to prove Theorem 2.1, it remains to check that

(i) all the conditions required in [18, Theorem 3.3, p.307] are satisfied, ensuring thus the existence of a solution y_*^h of (A.6) for all sufficiently small h; and that

(ii) the solution y_*^h of the first-order optimality condition (A.6) obtained from (i) is indeed a minimizer of (6).

We proceed in two steps below.

Step 1. Introducing the space $Y := L^2(0, T; \mathbb{R}^d) \times \mathbb{R}^d$, the conditions required in [18, Theorem 3.3, p.307] are:

(C1) y_* is a nonsingular solution to (A.5) in the sense that (A.18) holds.
(C2) $\mathcal{G} : X \to Y$ is C^2 smooth, and $\mathrm{D}^2\mathcal{G}$ is bounded on all bounded subsets of X.
(C3) There exists a subspace Z of Y, with continuous embedding, such that $\mathrm{D}\mathcal{G}(u) \in L(X, Z)$, $\forall u \in X$.
(C4) $\lim_{h \to 0} \|(\mathcal{T} - \mathcal{T}^h)(f, g)\|_X = 0$, $\forall (f, g) \in Y$.
(C5) $\lim_{h \to 0} \|\mathcal{T} - \mathcal{T}^h\|_{L(Z, X)} = 0$.

Verification of (C1). To guarantee (A.18), it suffices to show that

$$\|\mathcal{T} \circ \mathrm{D}\mathcal{G}(y_*)\|_{L(X, X)} < 1. \tag{A.19}$$

To this end, for any given v in X, let us denote $w = \mathcal{T} \circ \mathrm{D}\mathcal{G}(y_*)v$. By a direct calculation using the definition of $\mathcal{T}$ and $\mathcal{G}$ given respectively in (A.1) and (A.3), we get

$$w(t) = -\int_0^t \mathrm{D}F(y_*(s))v(s)\,\mathrm{d}s$$
$$- \int_0^t \int_s^T [\mathrm{D}F(y_*(\tau))]^\top \Big(v'(\tau) - \mathrm{D}F(y_*(\tau))v(\tau)\Big)\,\mathrm{d}\tau\,\mathrm{d}s$$
$$- \int_0^T [\mathrm{D}F(y_*(s))]^\top \Big(v'(s) - \mathrm{D}F(y_*(s))v(s)\Big)\,\mathrm{d}s, \quad t \in [0, T].$$

Introducing

$$M := \sup_{s \in [0, T]} \|\mathrm{D}F(y_*(s))\|_{\mathrm{op}}^2, \tag{A.20}$$

we obtain by a direct estimation based on the Hölder's inequality that

$$\langle w(0), w(0) \rangle \leq 2TM\big(\|v'\|_{L^2(0, T; \mathbb{R}^d)}^2 + M\|v\|_{L^2(0, T; \mathbb{R}^d)}^2\big)$$
$$\leq 2TM\big(1 + \widetilde{\mathfrak{C}}^2 M\big)\|v\|_X^2, \tag{A.21}$$

and that

$$\int_0^T \langle w'(t), w'(t) \rangle\,\mathrm{d}t \leq M(2\widetilde{\mathfrak{C}}^2 + 4T^2 + 4T^2\widetilde{\mathfrak{C}}^2 M)\|v\|_X^2. \tag{A.22}$$

Recalling that $w = \mathcal{T} \circ \mathrm{D}\mathcal{G}(y_*)v$, we get from (A.21) and (A.22) that

$$\|\mathcal{T} \circ \mathrm{D}\mathcal{G}(y_*)v\|_X^2 \leq 2M(2T^2 + T + \widetilde{\mathfrak{C}}^2 + T\widetilde{\mathfrak{C}}^2(1 + 2T)M)\|v\|_X^2. \tag{A.23}$$

---

[2] The formulation of [18, Theorem 3.3, p.307] concerns actually a parameterized family of (A.5) in which the nonlinearity $\mathcal{G}$ depends on an additional scalar parameter λ. Since there is no such a parameter in our setting, it can be viewed as a special case for which λ is taken to be a constant here.

Since T is fixed, we get $\|\mathcal{T} \circ \mathrm{D}\mathcal{G}(y_*)\,v\|_X < \|v\|_X^2$ for all v in X when M satisfies

$$2M(2T^2 + T + \widetilde{\mathfrak{C}}^2 + T\widetilde{\mathfrak{C}}^2(1 + 2T)M) < 1.$$

That is

$$M = \sup_{s \in [0,T]} \|\mathrm{D}F(y_*(s))\|_{\mathrm{op}}^2$$

$$< \frac{1}{2T^2 + T + \widetilde{\mathfrak{C}}^2 + \sqrt{(2T^2 + T + \widetilde{\mathfrak{C}}^2)^2 + 2T\widetilde{\mathfrak{C}}^2(1 + 2T)}}. \tag{A.24}$$

We have thus verified (A.19) when $\sup_{s \in [0,T]} \|\mathrm{D}F(y_*(s))\|_{\mathrm{op}}$ is small such that (A.24) holds. Consequently, Condition **(C1)** holds under this smallness assumption on $\mathrm{D}F$.

Verification of (C2). This condition can be checked by a long but straightforward calculation using the explicit form of $\mathcal{G}$ given by (A.3)–(A.4) and the assumption that F is C^3 smooth.

Verification of (C3). We take $Z := H^1(0, T; \mathbb{R}^d) \times \mathbb{R}^d$. Note that Z is compactly embedded into Y. Condition **(C3)** follows then from a direct but lengthy calculation based again on the explicit form of $\mathcal{G}$ given by (A.3)–(A.4). It suffices to assume F to be C^2 smooth to check this condition.

Verification of (C4). Recall from (A.17) that $\mathcal{T}^h = \Pi_h \mathcal{T}$. Condition **(C4)** follows immediately because $\|(\mathrm{Id}_X - \Pi_h)u\|_X$ converges to zero as h goes to zero for all $u \in X$ as a property of the finite element subspaces X^h.

Verification of (C5). As pointed out in [18, Theorem 3.3, p.307], Condition **(C5)** is a consequence of Condition **(C4)** (and the uniform boundedness theorem) when Z is compactly embedded into Y, which is the case here for $Z = H^1(0, T; \mathbb{R}^d) \times \mathbb{R}^d$. See also [12, Lemma 8.7].

All the conditions in [18, Theorem 3.3, p.307] are thus verified. It follows then from this theorem that for any given neighborhood $\mathcal{O}$ of y_*, the problem (6) has a unique solution y_*^h in $\mathcal{O}$ for all sufficiently small h; and the convergence result (12) holds.

Step 2. It remains to show that y_*^h obtained from Step 1 above is indeed a minimizer of (6). For this purpose, it suffices to show that

$$\frac{\mathrm{d}^2}{\mathrm{d}\tau^2} J(y_*^h + \tau v; F, f, g)|_{\tau=0} > 0, \quad \forall\, v \in X^h \backslash \{0\}. \tag{A.25}$$

Note that

$$\frac{\mathrm{d}^2}{\mathrm{d}\tau^2} J(y_*^h + \tau v; F, f, g)\Big|_{\tau=0}$$

$$= \int_0^T \langle -\mathrm{D}^2 F(y_*^h(t))(v(t), v(t)), (y_*^h)'(t) - F(y_*^h(t)) - f(t)\rangle\, \mathrm{d}t$$

$$+ \int_0^T \langle v'(t) - \mathrm{D}F(y_*^h(t))v(t), v'(t) - \mathrm{D}F(y_*^h(t))v(t)\rangle\, \mathrm{d}t + \langle v(0), v(0)\rangle,$$

where $\mathrm{D}^2 F(y_*^h(t))$ denotes the Hessian of F evaluated at $y_*^h(t)$, which is a bilinear function mapping $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathbb{R}^d$.

A straightforward estimation leads then to

$$\frac{\mathrm{d}^2}{\mathrm{d}\tau^2} J(y_*^h + \tau v; F, f, g)\Big|_{\tau=0} \geq \|v\|_X^2 \left(1 - 2\widetilde{\mathfrak{C}} \sup_{t \in [0,T]} \|\mathrm{D}F(y_*^h(t))\|_{\mathrm{op}} \right.$$

$$\left. - \mathfrak{C}\widetilde{\mathfrak{C}}\left(\sup_{t \in [0,T]} \left\|\!\left\|\mathrm{D}^2 F(y_*^h(t))\right\|\!\right\| \right) \|(y_*^h)' - F(y_*^h) - f\|_{L^2(0,T;\mathbb{R}^d)} \right), \tag{A.26}$$

where $\left\|\!\left\|\mathrm{D}^2 F(y_*(t))\right\|\!\right\|$ denotes the operator norm of the bilinear map $\mathrm{D}^2 F(y_*(t))$, $\mathfrak{C}$ and $\widetilde{\mathfrak{C}}$ are the embedding constants defined at the end of Section 2.1.

Since it has been shown in Step 1 that y_*^h converges in X-norm to y_* (cf. (12)) and X is continuously embedded into $C([0, T]; \mathbb{R}^d)$, we get

$$\lim_{h \to 0} \max_{t \in [0,T]} \|y_*(t) - y_*^h(t)\| = 0. \tag{A.27}$$

It follows that

$$\lim_{h \to 0} \sup_{t \in [0,T]} \|\mathrm{D}F(y_*^h(t))\|_{\mathrm{op}} = \sup_{t \in [0,T]} \|\mathrm{D}F(y_*(t))\|_{\mathrm{op}}. \tag{A.28}$$

The convergence results (12) and (A.27) together with the smoothness of F also imply that

$$\lim_{h \to 0} \|(y_*^h)' - F(y_*^h) - f\|_{L^2(0,T;\mathbb{R}^d)} = \|(y_*)' - F(y_*) - f\|_{L^2(0,T;\mathbb{R}^d)} = 0, \tag{A.29}$$

where the second equality holds since y_* is a strong solution of the IVP (3).

Thanks also to (A.27), we know that $\sup_{t\in[0,T]} \left\| D^2 F(y_*^h(t)) \right\|$ is uniformly bounded with respect to h. This uniform boundedness together with (A.28) and (A.29) implies that for all nonzero v in X^h, the right-hand side of (A.26) is positive when h is sufficiently small provided that

$$\sup_{t\in[0,T]} \|DF(y_*(t))\|_{\text{op}} < \frac{1}{2\widetilde{\mathfrak{C}}}. \tag{A.30}$$

We have thus verified (A.25) under the condition (A.30) by taking h sufficiently small. The proof is now complete. □

Appendix B. Proofs of Lemma 3.1

Note that (21) always has a solution since X^h is finite dimensional and the objective function is bounded below by zero. The fact $\widetilde{y}_*^h = \Pi_h \widetilde{y}_*$ is the unique solution to (21) follows directly by inspecting the associated first-order optimality condition. This condition can be obtained from (A.8) by setting F to zero and restricting v to X^h, and it reads as follows

$$\int_0^T \langle (\widetilde{y}_*^h)' - f, (v^h)' \rangle \, \mathrm{d}t + \langle \widetilde{y}_*^h(0) - g, v^h(0) \rangle = 0, \qquad \forall \, v^h \in X^h. \tag{B.1}$$

Using the expression of the solution $\widetilde{y}_*$ given by (20) and the definition of the inner product $\langle \cdot, \cdot \rangle_X$ given by (7), we can rewrite the above condition as

$$\langle \widetilde{y}_*^h - \widetilde{y}_*, v^h \rangle_X = 0, \qquad \forall \, v^h \in X^h. \tag{B.2}$$

Hence, $\widetilde{y}_*^h - \widetilde{y}_*$ lives in the orthogonal complement of X^h. We get thus, $\widetilde{y}_*^h = \Pi_h \widetilde{y}_*$.

For the error estimate (22), see e.g., [14, Section 2.7.3] for a proof that relies on the classical Aubin–Nitsche trick. The proof presented therein deals with the special case $g = 0$ and for state space dimension $d = 1$. For $d > 1$, since the vector field is independent of the unknown variable, we can carry out the estimate component by component, which reduces the problem to the case $d = 1$. The general case of $g \neq 0$ can be handled by considering $z = \widetilde{y} - g$. □

References

[1] U.M. Ascher, R.M.M. Mattheij, R.D. Russell, Numerical Solution of Boundary Value Problems for Ordinary Differential Equations, in: Classics in Applied Mathematics, vol. 13, SIAM, Philadelphia, PA, 1995.

[2] E. Hairer, C. Lubich, G. Wanner, Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations, 2nd, in: Springer Series in Computational Mathematics, vol. 31, Springer, New York, 2006.

[3] G. Hairer, S.P. Nørsett, E. Wanner, Solving Ordinary Differential Equations I: Nonstiff Problems, 2nd edn., Springer, New York, 1993.

[4] G. Hairer, E. Wanner, Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, 2nd edn., Springer, New York, 1996.

[5] J.D. Lambert, Numerical Methods for Ordinary Differential Systems: The Initial Value Problem, Wiley New York, 1991.

[6] B. Cockburn, G.E. Karniadakis, C.-W. Shu, The development of discontinuous Galerkin methods, in: Discontinuous Galerkin Methods, in: Lect. Notes Comput. Sci. Eng., vol. 11, Springer, Berlin, 2000, pp. 3–50.

[7] G. Fairweather, D. Meade, A survey of spline collocation methods for the numerical solution of differential equations, in: Mathematics for Large Scale Computing, in: Lecture Notes in Pure and Appl. Math., vol. 120, Dekker, New York, 1989, pp. 297–341.

[8] A. Bellen, M. Zennaro, Numerical Methods for Delay Differential Equations, Oxford University Press, Oxford, 2013.

[9] E. Haber, L. Ruthotto, Stable architectures for deep neural networks, Inverse Problems 34 (1) (2017) 014004.

[10] S. Günther, W. Pazner, D. Qi, Spline parameterization of neural network controls for deep learning, 2021, arXiv preprint arXiv:2103.00301.

[11] E.B. Becker, G.F. Carey, J.T. Oden, Finite Elements, an Introduction: Volume I, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1981.

[12] P.B. Bochev, M.D. Gunzburger, Least-Squares Finite Element Methods, Springer, New York, 2009.

[13] M.E. Davis, Numerical Methods and Modeling for Chemical Engineers, John Wiley & Sons, Inc., New York, 1984.

[14] B. Jiang, The Least-Squares Finite Element Method: Theory and Applications in Computational Fluid Dynamics and Electromagnetics, Springer, New York, 1998.

[15] J.P. Aubin, Behavior of the error of the approximate solutions of boundary value problems for linear elliptic operators by Galerkin's and finite difference methods, Ann. Della Scuola Normale Superiore Di Pisa, Classe Di Scienze 21 (1967) 599–637.

[16] P.G. Ciarlet, The Finite Element Method for Elliptic Problems, in: Classics in Applied Mathematics, vol. 40, SIAM, Philadelphia, 2002.

[17] J. Nitsche, Ein Kriterium für die Quasi-Optimalität des Ritzschen Verfahrens, Numer. Math. 11 (1968) 346–348.

[18] V. Girault, P.-A. Raviart, Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms, Springer, New York, 1986.

[19] H. Brézis, Functional Analysis, Sobolev Spaces and Partial Differential Equations, Springer, New York, 2010.

[20] C. de Boor, A Practical Guide to Splines, 1st edn., Springer-Verlag, New York, NY, 2001.

[21] L.L. Schumaker, Spline Functions: Computational Methods, SIAM, Philadelphia, PA, 2015.

[22] P. Rentrop, An algorithm for the computation of the exponential spline, Numer. Math. 35 (1) (1980) 81–93.

[23] E. Kreyszig, H. Kreyszig, E.J. Norminton, Advanced Engineering Mathematics, 10th edn., Wiley, Hoboken, NJ, 2011.

[24] G.H. Golub, J.H. Welsch, Calculation of Gauss Quadrature Rules, Technical report, Stanford University, Stanford, CA, USA, 1967.

[25] A. Griewank, A. Walther, Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, Vol. 105, SIAM, 2008.

[26] R. Horst, P.M. Pardalos, Handbook of Global Optimization, Vol. 2, Springer, 2013.

[27] L.A. Segel, On the validity of the steady state assumption of enzyme kinetics, Bull. Math. Biol. 50 (6) (1988) 579–593.

[28] J. Eilertsen, S. Schnell, The quasi-steady-state approximations revisited: Timescales, small parameters, singularities, and normal forms in enzyme kinetics, Math. Biosci. 325 (2020) 108339.

[29] L. Perko, Differential Equations and Dynamical Systems, in: Texts in Applied Mathematics, vol. 7, Springer-Verlag, New York, 1991, p. xii+403.

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

[31] A. Ilchmann, T. Reis, Surveys in Differential-Algebraic Equations II, Springer, 2014.

[32] M. Chung, M. Binois, R.B. Gramacy, J.M. Bardsley, D.J. Moquin, A.P. Smith, A.M. Smith, Parameter and uncertainty estimation for dynamical systems using surrogate stochastic processes, SIAM J. Sci. Comput. 41 (4) (2019) A2212–A2238.

[33] M. Chung, J. Krueger, M. Pop, Identification of microbiota dynamics using robust parameter estimation methods, Math. Biosci. 294 (2017) 71–84.

[34] M. Chung, E. Haber, Experimental design for biological systems, SIAM J. Control Optim. 50 (1) (2012) 471–489.

[35] A. Carracedo Rodriguez, M. Chung, S.M. Ciupe, Understanding the complex patterns observed during hepatitis b virus therapy, Viruses 9 (5) (2017) 117.

[36] L.C. Evans, Partial Differential Equations, in: Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, RI, 2010 p. xxii+749.

[37] H.L. Royden, P.M. Fitzpatrick, Real Analysis, 4th edn., Pearson, Boston, MA, 2010.