Machine learning of continuous and discrete variational ODEs with convergence guarantee and uncertainty quantification

Christian Offen

Paderborn University, Department of Mathematics

Warburger Str. 100, 33098 Paderborn, Germany

christian.offen@uni-paderborn.de

February 18, 2025

The article introduces a method to learn dynamical systems that are governed by Euler–Lagrange equations from data. The method is based on Gaussian process regression and identifies continuous or discrete Lagrangians and is, therefore, structure preserving by design. A rigorous proof of convergence as the distance between observation data points converges to zero is provided. Next to convergence guarantees, the method allows for quantification of model uncertainty, which can provide a basis of adaptive sampling techniques. We provide efficient uncertainty quantification of any observable that is linear in the Lagrangian, including of Hamiltonian functions (energy) and symplectic structures, which is of interest in the context of system identification. The article overcomes major practical and theoretical difficulties related to the ill-posedness of the identification task of (discrete) Lagrangians through a careful design of geometric regularisation strategies and through an exploit of a relation to convex minimisation problems in reproducing kernel Hilbert spaces.

1. Introduction

The identification of models of dynamical systems from data is an important task in machine learning with applications in engineering, physics, and molecular biology. Data-driven models are required when explicit descriptions for the equations of motions of dynamical systems are either not known or analytic descriptions are too computationally complex for large scale simulations. This contribution focuses on structure-preserving machine learning of dynamical systems based on Gaussian process regression and Gaussian fields. The framework allows for a rigorous convergence analysis and numerically efficient uncertainty estimation. The proposed method is a Lagrangian-based data-driven model. Let us briefly contrast the approach to Hamiltonian data-driven models and other Lagrangian-based models.

Hamiltonian data-driven models Physics-based, data-driven modelling aims to exploit prior physical or geometric knowledge when developing data-driven surrogate models of dynamical systems. Recent activities have developed methods to learn Hamiltonian systems, i.e. systems of the form

$$\dot{z} = J^{-1}\nabla H(z),\ J = \begin{pmatrix} 0_{d\times d} & -1_{d\times d} \\ 1_{d\times d} & 0_{d\times d} \end{pmatrix} \quad H\colon \mathbb{R}^{2d} \to \mathbb{R}^{2d}\,\text{(Hamiltonian)},$$

or port-Hamiltonian systems from data by approximating the Hamiltonian, pseudo-, or port-Hamiltonian structure by neural networks or Gaussian processes [25, 21, 7, 43, 41, 17]. Additionally, Lie group symmetries are identified in [18]. Alternatively, the symplectic flow map of Hamiltonian systems can be approximated [50, 11, 29]. The data-driven identification of interaction-based agent systems in [23, 31] or general Hamiltonian systems in [28] employ similar statistical learning methods as in this article but in the context of Hamiltonian systems. In contrast to the variational models considered in this article, Hamiltonian data-driven models mostly require prior knowledge of the symplectic phase space structure and observations of position and momenta, while the proposed Lagrangian-based methods only require observations of positions. Symplectic structures and Hamiltonians, however, can be derived from a Lagrangian model in a post-processing step. Approaches based on identifying symplectic structures or canonical coordinates from data together with a Hamiltonian have been considered, for instance, in [7, 13]. However, these do not provide a systematic discussion of uncertainty quantification or regularisation of this ill-posed inverse problem.

Continuous Lagrangian data-driven models Similarly to Hamiltonian data-driven models, variational principles for dynamical systems have been identified from data by identifying a Lagrangian function of the system [16, 37, 22, 30]. We refer to [34, 4] for an introduction to Lagrangian mechanics. To recall briefly, a dynamical system is governed by a *variational principle* or a *least action principle*, if motions constitute critical points of an action functional. In case of an autonomous first-order time-dependent system, the action functional is of the form

$$S(x) = \int_{t_0}^{t_1} L(x(t), \dot{x}(t))\mathrm{d}t, \tag{1}$$

where $x\colon [t_0, t_1] \to \mathbb{R}^d$ is a curve with derivative denoted by $\dot{x}$. The function L is a *Lagrangian.* A function $x\colon [t_0, t_1] \to \mathbb{R}^d$ is a solution or *motion* if the action S is stationary at x for all variations $\delta x\colon [t_0, t_1] \to \mathbb{R}^d$ that fix the endpoints t_0, t_1. Regularity assumptions on L and x provided, this is equivalent to the condition that x fulfils the Euler-Lagrange equations

$$\mathrm{EL}(L)(x(t), \dot{x}(t), \ddot{x}(t)) = 0, \qquad t \in (t_0, t_1) \tag{2}$$

with

$$\mathrm{EL}(L) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = \frac{\partial^2 L}{\partial \dot{x}\partial \dot{x}}\ddot{x} + \frac{\partial^2 L}{\partial \dot{x}\partial x}\dot{x} - \frac{\partial L}{\partial x}. \tag{3}$$

Here, $\frac{\partial^2 L}{\partial \dot{x}\partial \dot{x}} = \left(\frac{\partial^2 L}{\partial \dot{x}^k \partial \dot{x}^l}\right)_{k,l=1}^d$, $\frac{\partial^2 L}{\partial \dot{x}\partial x} = \left(\frac{\partial^2 L}{\partial \dot{x}^k \partial x^l}\right)_{k,l=1}^d$ refer to $d \times d$-dimensional blocks of the Hessian of L and $\frac{\partial L}{\partial x}$ denotes the gradient. Details may be found in [24, 51], for instance.

In the data-driven context, L is sought as a function of $\overline{x} = (x, \dot{x})$ such that (3) is fulfilled at observed data points $\mathcal{D} = \{(x, \dot{x}, \ddot{x})\}_{j=1}^M$. Once L is known, (2) can be solved with a numerical method such as a variational integrator [35].

Discrete Lagrangian data-driven models Instead of learning continuous variational principles, in [46] Qin proposes to learn discrete Lagrangian theories by approximating discrete Lagrangians. In discrete Lagrangian theories, motions $x(t)$ are described at discrete, equidistant times $t^0 < t^1 < \ldots < t^N$ by a sequence of snapshots $\boldsymbol{x} = (x_k)_{k=0}^N \subset \mathbb{R}^d$. The motions constitute stationary points of a discrete action functional

$$S_d(\boldsymbol{x}) = \sum_{k=1}^{N} L_d(x_{k-1}, x_k)$$

with respect to discrete variations of the interior points $x_1, \ldots, x_{N-1}$. In other words, $\boldsymbol{x}$ is a solution of the discrete field theory if $\frac{\partial S_d}{\partial x_k}(\boldsymbol{x}) = 0$ for all $1 \leq k < N$. This is equivalent to the discrete Euler–Lagrange equation

$$\mathrm{DEL}(L_d)(x_{k-1}, x_k, x_{k+1}) = 0, \quad 1 \leq k < N \tag{4}$$

with

$$\mathrm{DEL}(L_d)(x_{k-1}, x_k, x_{k+1}) = \nabla_2 L_d(x_{k-1}, x_k) + \nabla_1 L_d(x_k, x_{k+1}). \tag{5}$$

Here $\nabla_1 L_d$ and $\nabla_2 L_d$ denote the partial derivatives with respect to the first or second component of L_d, respectively. Details on discrete mechanics can be found in [35].

For the identification of discrete Lagrangians from data, training data $\mathcal{D} = \{x(t^k)\}_k$ consists of snapshots of motions of the dynamical system at discrete time-steps t^k. This needs to be contrasted to training of continuous Lagrangians which requires observations of first and second order derivatives of solutions, i.e. data of the form $\hat{x} = (x, \dot{x}, \ddot{x})$.

The class of discrete Lagrangian systems is expressive enough to describe motions of continuous Lagrangian systems on bounded open subsets of $\mathbb{R}^d$ at the snapshot times $(t^k)_k$ exactly, i.e. without discretisation error, provided the step-size $\Delta t = t^{k+1} - t^k$ is small enough, see [35, §1.6]. Thus, identifying L_d instead of L is fully justified from a modelling viewpoint. In case a continuous Lagrangian is required for system identification tasks or highly accurate predictions of velocity data, in the article [37] the author provides a method based on Vermeeren's variational backward error analysis [56] to recover continuous Lagrangians from data-driven discrete Lagrangians as a power series in the step-size of the time-grid.

Ambiguity of Lagrangians The data-driven identification of a continuous or discrete Lagrangian density is an ill-defined inverse problem as many different Lagrangian densities can yield equations of motions with the same set of solutions. This provides a challenge in a machine learning context and can lead to badly conditioned identified models that amplify errors [37]. In [42, 40] the author develops regularisation strategies that optimise numerical conditioning of the learnt theory, when the Lagrangian density is modelled as a neural network. The present article relates to Gaussian fields to allow for efficient uncertainty quantification and a theoretical convergence analysis.

Novelty The article

1. introduces a method to learn continuous and discrete Lagrangians from data based on Gaussian process regression with a rigorous convergence analysis as the distance between data points converges to zero.

2. Moreover, the article systematically discusses the ambiguity or Lagrangians and regularisation strategies for kernel-based learning methods for Lagrangians.

3. Furthermore, the article provides a statistical framework that allows for efficient uncertainty quantification of any linear observable of the dynamical system, such as Hamiltonian functions (energy) or symplectic structure, for instance. The uncertainty quantification does not require sampling but only to solve linear systems of equations.

This needs to be contrasted to aforementioned methods of the literature for learning Lagrangians, for which convergence guarantees are not provided or which do not provide uncertainty quantification of linear observables. Moreover, in the literature discussions on removing ambiguity of Lagrangians in data-driven identification are mostly absent: its necessity is sometimes avoided by assuming that torques are observed [22], an explicit mechanical ansatz is used [2]. In other works regularisation is done implicitly without discussion [16], ad hoc as in the author's prior work [37], or relates to neural networks [30, 42, 40] only.

Methodologically, the method of the present article stands in the context of meshless collocation methods [52] for solving linear partial differential equations since it solves (3) for L. It overcomes the major technical difficulty to prove convergence even though the Lagrangian density is not unique even after regularisation. For this, the article exploits a relation between posterior means of Gaussian processes and constraint optimisation problems in reproducing kernel Hilbert spaces that was presented in a game theory context by Owhadi and Scovel in [44] and was employed to solve well-posed partial differential equations using Gaussian Processes in [12].

Outline The article proceeds as follows: Section 2 continues the review of continuous and discrete variational principles that was started in the introduction. Moreover, it presents symplectic structure and Hamiltonians as linear observables of Lagrangian systems and it reviews the ambiguity of Lagrangians. Section 3 introduces methods to regularise the inverse problem of finding Lagrangian densities given dynamical data. In Section 4 we briefly review reproducing kernel Hilbert spaces and aspects of Gaussian fields. A more detailed discussion of the underlying theoretical concepts is provided in Appendix A. The section proceeds with an introduction of our method to learn continuous and discrete Lagrangians and to provide uncertainty quantifications for linear observables. Section 5 contains numerical experiments including identification of a Lagrangian and Hamiltonian for the coupled harmonic oscillator and convergence tests. Section 6 provides a theoretical convergence analysis of the method including a proof of the method's convergence. Additionally, convergence rates are derived. The article concludes with a summary in Section 7.

2. Background - Lagrangian dynamics

2.1. Continuous Lagrangian theories

2.1.1. Definition of associated Hamiltonian and symplectic structure

Let us continue our review of Lagrangian dynamics to fix notations and to explain the ambiguity that is inherent in the inverse problem of identifying (discrete) Lagrangians to observed motions. We postpone a provision of a more detailed functional analytic settings to the convergence analysis of Section 6 and refer to the literature on variational calculus [24, 51] for details.

We consider the Hamiltonian to a Lagrangian defined via

$$\mathrm{Ham}(L)(x,\dot{x}) = \dot{x}^\top \frac{\partial L}{\partial \dot{x}}(x,\dot{x}) - L(x,\dot{x}). \tag{6}$$

Here $\dot{x}^\top$ denotes the transpose of $\dot{x} \in \mathbb{R}^d$. The Hamiltonian $\mathrm{Ham}(L)$ is conserved along solutions of (2). Moreover, we consider the symplectic structure related to L which is given as the closed differential 2-form

$$\mathrm{Sympl}(L) = \sum_{s=1}^{d} \mathrm{d}x^s \wedge \mathrm{d}\left(\frac{\partial L}{\partial \dot{x}^s}\right) = \sum_{s,r=1}^{d} \frac{\partial^2 L}{\partial x^r \partial \dot{x}^s} \mathrm{d}x^s \wedge \mathrm{d}x^r + \frac{\partial^2 L}{\partial \dot{x}^r \partial \dot{x}^s} \mathrm{d}x^s \wedge \mathrm{d}\dot{x}^r. \tag{7}$$

When $\frac{\partial^2 L}{\partial \dot{x} \partial \dot{x}}$ is invertible everywhere, then the differential form $\mathrm{Sympl}(L)$ is non-degenerate and, therefore, a symplectic form.[1] As an aside, the motions (2) can be described as Hamiltonian motions to the Hamiltonian $\mathrm{Ham}(L)$ and symplectic structure $\mathrm{Sympl}(L)$. Moreover, we consider the induced momenta

$$\mathrm{Mm}(L)(x,\dot{x}) = \frac{\partial L}{\partial \dot{x}}(x,\dot{x}). \tag{8}$$

[1] $\mathrm{Sympl}(L)$ is the pull-back of the canonical symplectic form $\sum_{s=1}^d \mathrm{d}q^s \wedge \mathrm{d}p_s$ under the Legendre transform $T\mathbb{R}^d \to T^*\mathbb{R}^d$, $(x,\dot{x}) \mapsto (q,p) = (x, \frac{\partial L}{\partial \dot{x}}(x,\dot{x}))$.

Additionally, we consider the induced Liouville volume form given as the dth exterior power of $\mathrm{Sympl}(L)$

$$\mathrm{Vol}(L) = \frac{1}{d!}(\mathrm{Sympl}(L))^d = \det\left(\frac{\partial^2 L}{\partial \dot{x}^r \partial \dot{x}^s}\right) \mathrm{d}x^1 \wedge \mathrm{d}\dot{x}^1 \wedge \ldots \wedge \mathrm{d}x^d \wedge \mathrm{d}\dot{x}^d. \tag{9}$$

It will be of significance later that EL, Ham, Sympl, Mm are linear in the Lagrangian L, while Vol is not.

Example 1 Consider a mechanical Lagrangian $L(x,\dot{x}) = \frac{1}{2}\dot{x}^\top M\dot{x} - V(x)$ for a continuously differentiable potential $V\colon \mathbb{R}^d \to \mathbb{R}$ and a symmetric, positive definite matrix M (mass matrix). The equations of motions are $0 = \mathrm{EL}(L)(x,\dot{x},\ddot{x}) = \ddot{x} + \nabla V(x)$, where $\nabla V = \frac{\partial V}{\partial x}$ denotes the gradient of V. The conjugate momentum is $p := \mathrm{Mm}(L)(x,\dot{x}) = M\dot{x}$. The Hamiltonian function is $H(x,p) = \mathrm{Ham}(L)(x, M^{-1}p) = \frac{1}{2}p^\top M^{-1}p + V(x)$. The symplectic form is $\omega = \mathrm{Symp}(L) = \sum_{s=1}^d \mathrm{d}x^s \wedge \mathrm{d}p^s$. In the frame induced by the coordinates (x,p) of the phase space the symplectic form is represented by the block matrix

$$J = \begin{pmatrix} 0_{n\times n} & -1_{n\times n} \\ 1_{n\times n} & 0_{n\times n} \end{pmatrix}.$$

Here $0_{n\times n}$ and $1_{n\times n}$ denote the zero and the identity matrix of size $n \times n$, respectively. In the coordinates (x,p), the equations of motions are Hamilton's equations in their standard form

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = J^{-1}\nabla H(x,p) = \begin{pmatrix} M^{-1}p \\ -\nabla V(x) \end{pmatrix}$$

The volume form $\mathrm{Vol}(L) = \det(M)\mathrm{d}x^1 \wedge \mathrm{d}\dot{x}^1 \wedge \ldots \wedge \mathrm{d}x^d \wedge \mathrm{d}\dot{x}^d = \mathrm{d}x^1 \wedge \mathrm{d}p^1 \wedge \ldots \wedge \mathrm{d}x^d \wedge \mathrm{d}p^d$ is the standard Euclidean volume form on the phase space. □

2.1.2. Ambiguity of Lagrangian densities

The ambiguity of Lagrangians in the description of variational dynamical systems has been the subject of various articles in theoretical physics including [27, 33, 32]. Lagrangians can be ambiguous in two different ways:

1. Lagrangians L and $\tilde{L}$ can yield the same Euler–Lagrange operator (3) up to rescaling, i.e.

   $$\rho\mathrm{EL}(L) = \mathrm{EL}(\tilde{L}), \quad \rho \in \mathbb{R}\setminus\{0\}$$

   and, therefore, the same Euler–Lagrange equations (2) up to rescaling. We call L and $\tilde{L}$ *(gauge-) equivalent*. For equivalent Lagrangians L, $\tilde{L}$ there exists $\rho \in \mathbb{R}\setminus\{0\}$, $c \in \mathbb{R}$ such that $\tilde{L} - \rho L - c$ is a total derivative

   $$\tilde{L} - \rho L - c = \mathrm{d}_t F$$

for a continuously differentiable function $F\colon \mathbb{R}^d \to \mathbb{R}$, where

$$\mathrm{d}_t F(x,\dot{x}) = \dot{x}^\top \nabla F(x) = \sum_{s=1}^{d} \dot{x}^s \frac{\partial F}{\partial x^s}(x) \tag{10}$$

(See, e.g. [24].) We have restricted ourselves to autonomous Lagrangians.

2. More generally, two Lagrangians L and $\tilde{L}$ can yield the same set of solutions x, i.e.

$$\mathrm{EL}(L)(x(t),\dot{x}(t)),\ddot{x}(t)) = 0 \iff \mathrm{EL}(\tilde{L})(x(t),\dot{x}(t)),\ddot{x}(t)) = 0$$

for all regular curves $x\colon [t_0,t_1] \to \mathbb{R}^d$ even when they are *not* equivalent in the sense of Item 1. In such a case, $\tilde{L}$ is called an *alternative Lagrangian* to L.

Example 2 (Affine linear motions) For any twice differentiable $g\colon \mathbb{R}^d \to \mathbb{R}$ with nowhere degenerate Hessian matrix $\mathrm{Hess}(g)$, the Lagrangian $L(x,\dot{x}) = g(\dot{x})$ describes affine linear motions in $\mathbb{R}^d$:

$$0 = \mathrm{EL}(L) = \mathrm{Hess}(g)(\dot{x})\ddot{x}.$$

□

In general, the existence of alternative Lagrangian densities is related to additional geometric structure and conserved quantities of the system [27, 33, 32, 10]. This article mainly considers ambiguities by equivalence, which are exhibited by all variational systems.

Lemma 1 *Let L be a Lagrangian depending on $(x,\dot{x})$. Consider a continuously differentiable $F\colon \mathbb{R}^d \to \mathbb{R}$, $\rho \in \mathbb{R}$, $c \in \mathbb{R}$, and $\tilde{L} = \rho L + \mathrm{d}_t F + c$. We have*

$$\begin{aligned}
\mathrm{EL}(\tilde{L}) &= \rho \mathrm{EL}(L)\\
\mathrm{Mm}(\tilde{L}) &= \rho \mathrm{Mm}(L) + \nabla F\\
\mathrm{Sympl}(\tilde{L}) &= \rho \mathrm{Sympl}(L)\\
\mathrm{Vol}(\tilde{L}) &= \rho^d \mathrm{Vol}(L)\\
\mathrm{Ham}(\tilde{L}) &= \rho \mathrm{Ham}(L) - c
\end{aligned}$$

Here ∇F denotes the gradient of F. Moreover, if $\rho \neq 0$ then

$$\left\{(x,\dot{x}) : \det\left(\frac{\partial^2 L}{\partial \dot{x}\partial \dot{x}}\right)(x,\dot{x}) \neq 0\right\} = \left\{(x,\dot{x}) : \det\left(\frac{\partial^2 \tilde{L}}{\partial \dot{x}\partial \dot{x}}\right)(x,\dot{x}) \neq 0\right\}. \tag{11}$$

□

Proof The transformation rules of EL, Mm, Sympl, Vol, and Ham are obtained by a direct computation. The assertion (11) follows from the transformation rule for Vol or directly by observing that $\frac{\partial^2 \tilde{L}}{\partial \dot{x}\partial \dot{x}} = \rho \frac{\partial^2 L}{\partial \dot{x}\partial \dot{x}}$. ■

The following Corollary is a restatement of (11).

Corollary 1 *The set where a Lagrangian L is non-degenerate, i.e. where $\frac{\partial^2 L}{\partial \dot{x}\partial \dot{x}}$ is invertible, is invariant under equivalence.* □

2.2. Discrete Lagrangian systems

2.2.1. Associated symplectic structure

In analogy to the continuous case (Section 2.1.1) we define associated data to a discrete Lagrangian density $L_d\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ following definitions in discrete variational calculus [35]. The quantities

$$\begin{aligned}\mathrm{Mm}^-(L_d)(x_j, x_{j+1}) &= -\nabla_1 L_d(x_j, x_{j+1})\\ \mathrm{Mm}^+(L_d)(x_{j-1}, x_j) &= \nabla_2 L_d(x_{j-1}, x_j)\end{aligned}$$

relate to discrete conjugate momenta at time t_j. On motions $\boldsymbol{x} = (x_k)_{k=0}^N$ that fulfil (4), $\mathrm{Mm}^-(x_k, x_{k+1})$ and $\mathrm{Mm}^+(x_{k-1}, x_k)$ coincide for all $1 \leq k < N$. Moreover, denoting the coordinate of the domain of definition $\mathbb{R}^d \times \mathbb{R}^d$ of L_d by (x_0, x_1) we define the 2-form

$$\mathrm{Sympl}(L_d) = \sum_{r,s=1}^{d} \frac{\partial^2 L_d}{\partial x_1^s \partial x_0^r} \mathrm{d}x_1^s \wedge \mathrm{d}x_0^r \tag{12}$$

and its dth exterior power normalised by $\frac{1}{d!}$

$$\mathrm{Vol}(L_d) = \det\left(\frac{\partial^2 L_d}{\partial x_1 \partial x_0}\right) \mathrm{d}x_1^1 \wedge \mathrm{d}x_0^1 \wedge \ldots \wedge \mathrm{d}x_1^d \wedge \mathrm{d}x_0^d. \tag{13}$$

When $\frac{\partial^2 L_d}{\partial x_1 \partial x_0}$ is non-degenerate everywhere, then $\mathrm{Sympl}(L_d)$ is a symplectic form and $\mathrm{Vol}(L_d)$ its induced volume form on the discrete phase space $\mathbb{R}^d \times \mathbb{R}^d$. $\mathrm{Sympl}(L_d)$ is called *discrete Lagrangian symplectic form* in [35, §1.3.2]. (For consistency with the continuous theory Section 2.1.1 our sign convention differs from [35, §1.3.2]. A derivation can be found in Appendix C.)

2.3. Ambiguity of discrete Lagrangians

In analogy to Section 2.1.2, if L_d is a discrete Lagrangian and $\tilde{L}_d(x_0, x_1) = \rho L_d(x_0, x_1) + F(x_1) - F(x_0) + c$ for $c \in \mathbb{R}$, $\rho \in \mathbb{R} \setminus \{0\}$, and continuously differentiable F, then

$$\rho \mathrm{DEL}(L_d) = \mathrm{DEL}(\tilde{L}_d)$$

and L_d and $\tilde{L}_d$ are called *(gauge-) equivalent*. Non-equivalent discrete Lagrangians such that the discrete Euler–Lagrange equations (4) have the same solutions are called *alternative discrete Lagrangians*.

The analogy of Lemma 1 for discrete Lagrangians is as follows.

Lemma 2 *Let L_d be a discrete Lagrangian depending on (x_0, x_1). Consider a continuously differentiable $F\colon \mathbb{R}^d \to \mathbb{R}$, $\rho \in \mathbb{R}$, $c \in \mathbb{R}$, and $\tilde{L}_d = \rho L_d + \Delta_t F + c$ with*

$\Delta_t F(x_0,x_1) = F(x_1) - F(x_0)$. *We have*

$$\begin{aligned} \mathrm{DEL}(\tilde{L}_d) &= \rho \mathrm{DEL}(L_d) \\ \mathrm{Mm}^-(\tilde{L}_d)(x_0,x_1) &= \rho \mathrm{Mm}^-(L_d)(x_0,x_1) + \nabla F(x_0) \\ \mathrm{Mm}^+(\tilde{L}_d)(x_0,x_1) &= \rho \mathrm{Mm}^+(L_d)(x_0,x_1) + \nabla F(x_1) \\ \mathrm{Sympl}(\tilde{L}_d) &= \rho \mathrm{Sympl}(L_d) \\ \mathrm{Vol}(\tilde{L}_d) &= \rho^d \mathrm{Vol}(L_d) \end{aligned}$$

Here ∇F *denotes the gradient of* F. *Moreover, if* $\rho \neq 0$ *then*

$$\left\{(x_0,x_1) : \det\left(\frac{\partial^2 L_d}{\partial x_0 \partial x_1}\right)(x_0,x_1) \neq 0\right\} = \left\{(x_0,x_1) : \det\left(\frac{\partial^2 \tilde{L}_d}{\partial x_0 \partial x_1}\right)(x_0,x_1) \neq 0\right\}.$$

□

PROOF The transformation rules of EL, $\mathrm{Mm}^{\pm}$, Sympl, Vol are obtained by a direct computation. The assertion about invariance of non-degenerate points follows from the transformation rule of Vol. ■

3. Regularisation

In the machine learning framework that we will introduce in Section 4, we will employ regularisation conditions to safeguard us from finding degenerate solutions to the inverse problem of identifying a Lagrangian to given motions. Extreme instances of degenerate solutions are Null-Lagrangians, for which $\mathrm{EL}(L) \equiv 0$. These are consistent with any dynamics but cannot discriminate curves that are not motions.

The following section serves two goals:

- We justify that the employed regularisation conditions are covered by the ambiguities presented in Section 2. Therefore, imposing these on L does not restrict the generality of the ansatz. We will also refer to these as *normalisation conditions* as we will impose that these are fulfilled exactly by the data-driven model.
- The normalisation conditions (together with the system's motions) do *not* determine the Lagrangian uniquely. However, they guarantee that the sought Lagrangian is non-degenerate, provided that there are no true degenerate Lagrangians. Furthermore, we show that the normalisation conditions determine the symplectic structure $\mathrm{Sym}(L)$, the Hamiltonian $\mathrm{Ham}(L)$, and the Euler–Lagrange operator $\mathrm{EL}(L)$ of the system uniquely, provided that no true alternative Lagrangians exist. In the context of uncertainty quantification, this implies that any ambiguity in the representation of the model L does not contribute to uncertainty in the Hamiltonian, the symplectic structure, or the equations of motions. This justifies the approach towards uncertainty quantification in the article.

A reader mostly interested in the machine learning setting can skip ahead to Section 4.

3.1. Preparation of the regularisation strategy

Proposition 1 *Let* $\overline{x}_b = (x_b, \dot{x}_b) \in T\mathbb{R}^d \cong \mathbb{R}^d \times \mathbb{R}^d$*,* $\mathring{L}$ *a Lagrangian, and* $\hat{x}_\tau = (x_\tau, \dot{x}_\tau, \ddot{x}_\tau) \in (\mathbb{R}^d)^3$ *with* $\mathrm{EL}(\mathring{L})(\hat{x}_\tau) \neq 0$*.*[2] *Let* $c_b \in \mathbb{R}$*,* $p_b \in \mathbb{R}^d$*,* $c_\tau \neq 0$*. Then there exists a Lagrangian* L *such that* L *is equivalent to* $\mathring{L}$ *and*

$$L(\overline{x}_b) = c_b, \quad \mathrm{Mm}(L)(\overline{x}_b) = \frac{\partial L}{\partial \dot{x}}(\overline{x}_b) = p_b, \quad (\mathrm{EL}(L)(\hat{x}_\tau))_k = c_\tau, \tag{14}$$

where $1 \leq k \leq d$ *is any index for which the* k*th component of* $\mathrm{EL}(\mathring{L})(\hat{x}_\tau)$ *is not zero.* □

Proof Let $\mathring{c}_b = \mathring{L}(\overline{x}_b)$, $\mathring{p}_b = \mathrm{Mm}(\mathring{L})(\overline{x}_b)$, $\mathring{c}_\tau = (\mathrm{EL}(\mathring{L})(\hat{x}_\tau))_k$ (k th component). We set

$$\rho = \frac{c_\tau}{\mathring{c}_\tau}, \quad F(x) = x^\top (p_b - \rho \mathring{p}_b), \quad c = c_b - \dot{x}_b^\top (p_b - \rho \mathring{p}_b) - \rho \mathring{c}_b.$$

Now the Lagrangian $L = \rho \mathring{L} + \mathrm{d}_t F + c$ is equivalent to $\mathring{L}$ and fulfils (14). ■

While the equivalent Lagrangian L constructed in Proposition 1 is always non-degenerate if $\mathring{L}$ is non-degenerate (by Lemma 1), this is not necessarily true for all Lagrangians governing the motions even when restricting to those that fulfil (14): indeed, in Example 2 of affine linear motions governed by $\mathring{L}(x, \dot{x}) = \dot{x}^2$, we can choose g such that $L(x, \dot{x}) = g(\dot{x})$ has degenerate points at any points. However, when we exclude systems with alternative Lagrangians, then we have the following Proposition.

Proposition 2 *Let* $\mathring{L}$ *be a Lagrangian that is non-degenerate on some non-empty, connected set* $\mathcal{O} \subset T\mathbb{R}^d \cong \mathbb{R}^d \times \mathbb{R}^d$*. When no alternative Lagrangian to* $\mathring{L}$ *exists, then any Lagrangian* L *with the property*

$$\mathrm{EL}(\mathring{L})(x(t), \dot{x}(t), \ddot{x}(t)) = 0 \implies \mathrm{EL}(L)(x(t), \dot{x}(t), \ddot{x}(t)) = 0$$

on $\mathcal{O} \times \mathbb{R}^d$ *is either a null-Lagrangian (i.e.* $\mathrm{EL}(L) \equiv 0$*) or is non-degenerate on* $\mathcal{O}$*.* □

Proof As no alternative Lagrangian exists, there must be $\rho, c \in \mathbb{R}$ and $F \colon \mathbb{R}^d \to \mathbb{R}$ such that on $\mathcal{O}$

$$L = \rho \mathring{L} + \mathrm{d}_t F + c.$$

If L is not a null-Lagrangian on $\mathcal{O}$, there must be $\hat{x} \in \mathcal{O} \times \mathbb{R}^d$ with $\mathrm{EL}(L)(\hat{x}) \neq 0$. Let $1 \leq k \leq d$ such that $(\mathrm{EL}(L)(\hat{x}))_k \neq 0$. By Lemma 1

$$0 \neq (\mathrm{EL}(L)(\hat{x}))_k = \rho (\mathrm{EL}(\mathring{L})(\hat{x}))_k.$$

Thus $\rho \neq 0$. Non-degeneracy on $\mathcal{O}$ follows from $\mathrm{Vol}(L) = \rho^d \mathrm{Vol}(\mathring{L})$. ■

[2]This means that $\hat{x}_\tau = (x_\tau, \dot{x}_\tau, \ddot{x}_\tau)$ is any point that does not correspond to a motion of the dynamical system described by $\mathring{L}$. For instance, when $(x_\tau, \dot{x}_\tau)$ is an equilibrium point of the dynamics then we can chose any $\ddot{x}_\tau \neq 0$. The assumption excludes trivial Lagrangians such as $\mathring{L} \equiv 0$.

Remark 1 Under genericity assumptions on the dynamics with $d \geq 2$, no alternative Lagrangians exist [27]. If a generic dynamical system is governed by a non-degenerate Lagrangian, then any Lagrangian L with $\mathrm{EL}(L) = 0$ on all motions that is non-degenerate anywhere, is non-degenerate everywhere. □

Refer to Proposition 11 of Appendix B for an alternative normalisation strategy for Lagrangians based on normalising symplectic volume. It is comparable to techniques developed in [42] for neural network models of Lagrangians.

The following Proposition implies that the Euler–Lagrange operator (and thus the representation of the equation of motions) and the Hamiltonian and symplectic structure are uniquely determined when the normalisation condition (14) is fulfilled, provided that no alternative Lagrangians exist.

Proposition 3 *Let $\mathring{L}$ be a Lagrangian on $T\mathbb{R}^d$ with* (14) *for some $\overline{x}_b = (x_b, \dot{x}_b) \in T\mathbb{R}^d$, $1 \leq k \leq d$, $c_b \in \mathbb{R}$, $p_b \in \mathbb{R}^d$, $c_\tau \in \mathbb{R} \setminus \{0\}$. Then for any Lagrangian L with* (14) *that is equivalent to $\mathring{L}$ we have*

$$\mathrm{EL}(L) = \mathrm{EL}(\mathring{L}), \quad \mathrm{Ham}(L) = \mathrm{Ham}(\mathring{L}), \quad \mathrm{Sym}(L) = \mathrm{Sym}(\mathring{L}).$$ □

Proof L is of the form $L = \rho\mathring{L} + \mathrm{d}_t F + c$. The last condition of (14) implies $\rho = 1$. Thus $\mathrm{EL}(L) = \mathrm{EL}(\mathring{L})$ and $\mathrm{Sym}(L) = \mathrm{Sym}(\mathring{L})$ by Lemma 1. With $\rho = 1$ and the first two conditions (14) we have

$$\mathrm{Ham}(L)(\overline{x}_b) = \dot{x}_b^\top p_b - c_b = \mathrm{Ham}(\mathring{L})(\overline{x}_b).$$

Then $\mathrm{Ham}(L) = \mathrm{Ham}(\mathring{L})$ follows by Lemma 1. ■

For discrete Lagrangians, we have the following analogy to Proposition 1.

Proposition 4 *Let $\overline{x}_b = (x_{0b}, x_{1b}) \in (\mathbb{R}^d)^2$, $\hat{x}_\tau = (x_{0\tau}, x_{1\tau}, x_{2\tau}) \in (\mathbb{R}^d)^3$ and $\mathring{L}_d$ a discrete Lagrangian with $\mathrm{DEL}(L_d)(\hat{x}_b) \neq 0$. Let $c_b \in \mathbb{R}$, $p_b \in \mathbb{R}^d$, $c_\tau \in \mathbb{R} \setminus \{0\}$. There exists a discrete Lagrangian L_d such that L_d is equivalent to $\mathring{L}_d$ and*

$$L_d(\overline{x}_b) = c_b, \quad \mathrm{Mm}^+(L_d)(\overline{x}_b) = p_b, \quad (\mathrm{DEL}(L_d)(\hat{x}_\tau))_k = c_\tau, \tag{15}$$

where $1 \leq k \leq d$ can be chosen as any index for which the component of $\mathrm{DEL}(\hat{x}_b)$ is not zero. □

Proof Let $\mathring{c}_b = \mathring{L}_d(\overline{x}_b)$, $\mathring{p}_b = \mathrm{Mm}^+(\mathring{L}_d)(\overline{x}_b)$, $\mathring{c}_\tau = (\mathrm{DEL}(\mathring{L}_d)(\hat{x}_b))_k$. We set

$$\rho = \frac{c_\tau}{\mathring{c}_\tau}, \quad F(x) = x^\top(p_b - \rho\mathring{p}_b), \quad c = c_b - \rho\mathring{c}_b - (x_{1b} - x_{0b})^\top(p_b - \rho\mathring{p}_b).$$

Now the Lagrangian $L_d = \rho\mathring{L}_d + \Delta_t F + c$ is equivalent to L_d and fulfils (15). ■

Remark 2 A statement similar to Proposition 4 holds true with Mm^- replacing Mm^+. Moreover, a statement in analogy to Proposition 2 can be obtained with discrete quantities replacing their continuous counterparts. The details shall not be spelled out in this context. Moreover, an alternative normalisation strategy based on regularising the discrete symplectic volume is provided in Proposition 12 in Appendix B, where it is also compared to regularisation strategies in the neural network context of [42]. □

3.2. Utilisation in a data-driven context

In the following section, we will consider the inverse problem of inferring a Lagrangian or discrete Lagrangian from motion data. For this, we will augment the inverse problem by normalisation conditions (14) or (15), respectively, for values of $c_b \in \mathbb{R}$, $p_b \in \mathbb{R}^d$, and $c_\tau \in \mathbb{R} \setminus \{0\}$. Proposition 1 or Proposition 4 show that this augmentation does not restrict the generality of the ansatz. Although the conditions together with the true dynamics do not determine the (discrete) Lagrangian uniquely, they do determine the Euler–Lagrange operator $\mathrm{EL}(L)$ as well as the Hamiltonian and symplectic structure, provided that the true dynamical system does not admit alternative Lagrangians. When only limited data is observed, there is some uncertainty in the equations of motions $\mathrm{EL}(L) = 0$, the Hamiltonian, symplectic structure, or any linear observable in L that we want to quantify. The normalisation conditions eliminate any artificial uncertainty stemming from an ambiguous representation of the model.

Moreover, when all true Lagrangians are non-degenerate, so is the sought Lagrangian in the augmented inverse problem (Proposition 2). Thus, the normalisation conditions safeguard us from inferring degenerate Lagrangians that are consistent with the observed motion data but fail to discriminate non-motions.

4. Data-driven method

4.1. Bayesian learning of continuous Lagrangians

In the following, we present a framework for learning a continuous Lagrangian from observations of a dynamical system.

Let $\Omega \subset T\mathbb{R}^d \cong \mathbb{R}^d \times \mathbb{R}^d$ be an open, bounded subset. Our goal is to identify a Lagrangian $L \colon \Omega \to \mathbb{R}$ based on observations $\hat{x} = (\overline{x}, \ddot{x}) = (x, \dot{x}, \ddot{x}) \in \Omega \times \mathbb{R}^d$ for which $\mathrm{EL}(L)(\hat{x}) = 0$ on all observations $\hat{x}$ such that the dynamics (2) to L approximate the dynamics of an unknown true Lagrangian $L_{\mathrm{ref}} : \Omega \to \mathbb{R}$. We interpret this task as seeking a solution to the Euler–Lagrange equation (2) that we interpret as a partial differential equation for L. We follow a Bayesian approach proposed in [12] and assume a Gaussian field (see Appendix A for definitions) as a prior for L that we condition on fulfilling the Euler–Lagrange equation (2) on the data points and on regularisation conditions to obtain a posterior distribution for L. Even though in contrast to [12] our partial differential equation is highly ill-posed, we prove in Section 6 that the posterior mean converges against a true Lagrangian of the motions in the infinite data limit.

4.1.1. RKHS set-up and Gaussian fields

We consider the following set-up that makes use of the theory of reproducing kernel Hilbert spaces (RKHS). Refer to [14, 44] for background material.

Consider a symmetric function $K \colon \Omega \times \Omega \to \mathbb{R}$. Assume that K is positive definite, i.e. for all finite subsets $\{\overline{x}^{(j)}\}_{j=1}^M \subset \Omega$ the matrix $(K(\overline{x}^{(i)}, \overline{x}^{(j)}))_{i,j=1}^M$ is positive definite. ´K is called *kernel*.

Consider the reproducing kernel Hilbert space (RKHS) U to K, i.e. consider the inner product space

$$\mathring{U} = \left\{ L = \sum_{j=1}^{n} \alpha_j K(\overline{x}^{(j)}, \cdot) \mid \alpha_j \in \mathbb{R}, n \in \mathbb{N}_0, \overline{x}^{(j)} \in \Omega \right\}$$

with inner product defined as the linear extension of

$$\langle K(\overline{x}, \cdot), K(\overline{y}, \cdot) \rangle = K(\overline{x}, \overline{y}).$$

Then the Hilbert space U is obtained as the topological closure of $\mathring{U}$ with respect to $\langle \cdot, \cdot \rangle$. We denote the dual space of U by U^*. We define the map

$$\mathcal{K} \colon U^* \to U, \quad \Phi \mapsto \mathcal{K}(\Phi) \text{ with } \mathcal{K}(\Phi)(x) = \Phi(K(x, \cdot)). \tag{16}$$

The map $\mathcal{K} \colon U^* \to U$ is linear, bijective, and symmetric, i.e. $\Psi(\mathcal{K}(\Phi)) = \Phi(\mathcal{K}(\Psi))$ for $\Phi, \Psi \in U^*$, and positive, i.e. $\Phi(\mathcal{K}(\Phi)) > 0$ for $\Phi \in U^* \setminus \{0\}$.

Consider the *canonical Gaussian field*[3] $\xi \in \mathcal{N}(0, \mathcal{K})$ on U, which is a weak random variable with the following properties:

- For all $\phi \in U^*$, $\phi(\xi) \sim \mathcal{N}(0, \phi(\mathcal{K}(\phi)))$ is a centred Gaussian random variable.
- Moreover, for any finite collection $\Phi = (\Phi_1, \ldots, \Phi_n)$ with $\Phi_j \in U^*$ for $1 \leq j \leq n$, the random variable $\Phi(\xi)= (\Phi_1(\xi), \ldots, \Phi_n(\xi))$ is multivariate-normally distributed $\Phi(\xi) \in \mathcal{N}(0, \kappa)$ with covariance matrix given as $\kappa = (\Phi_i(\mathcal{K}(\Phi_j)))_{i,j=1}^{n}$.

See Appendix A for a formal definition of Gaussian fields and existence statements recalled from [44].

4.1.2. Data

Assume we observe distinct data points $\hat{x}^{(j)} = (\overline{x}^{(j)}, \ddot{x}^{(j)}) = (x^{(j)}, \dot{x}^{(j)}, \ddot{x}^{(j)}) \in \Omega \times \mathbb{R}^d$, $j = 1, \ldots, M$ of Lagrangian motions. Define $\mathrm{EL}_{\hat{x}^{(j)}} \colon U \to \mathbb{R}^d$ as

$$\mathrm{EL}_{\hat{x}^{(j)}}(L) = \mathrm{EL}(L)(\hat{x}^{(j)}) = \frac{\partial^2 L(\overline{x}^{(j)})}{\partial x \partial \dot{x}} \ddot{x}^{(j)} + \frac{\partial^2 L(\overline{x}^{(j)})}{\partial x \partial x} \dot{x}^{(j)} - \frac{\partial L(\overline{x}^{(j)})}{\partial x}$$

for $1 \leq j \leq M$. Furthermore, let $\overline{x}_b = (x_b, \dot{x}_b) \in \Omega$ and consider $\mathrm{Mm}_{\overline{x}_b} \colon U \to \mathbb{R}^d$ defined as

$$\mathrm{Mm}_{\overline{x}_b}(L) = \mathrm{Mm}(L)(\overline{x}_b) = \frac{\partial L}{\partial \dot{x}}(\overline{x}_b).$$

Moreover, let $\mathrm{ev}_{\overline{x}_b} \colon U \to \mathbb{R}$ with

$$\mathrm{ev}_{\overline{x}_b}(L) = L(\overline{x}_b)$$

[3] The notion of a *Gaussian field* differs slightly from the notion of a *Gaussian process* [15, Def.3]. See [45, §3.5-§4, paragraph 1] for further explanation. However, the literature refers to methods that solve pdes using the concept of Gaussian fields as *Gaussian processes based* methods (e.g. [12, 6]).

denote the evaluation functional. Collect these functionals in a linear map $\Phi_b^M : U \to (\mathbb{R}^d)^M \times \mathbb{R}^d \times \mathbb{R}$

$$\Phi_b^M = (\mathrm{EL}_{\hat{x}^{(1)}}, \ldots, \mathrm{EL}_{\hat{x}^{(M)}}, \mathrm{Mm}_{\overline{x}_b}, \mathrm{ev}_{\overline{x}_b}). \tag{17}$$

For constants $c_b \in \mathbb{R}$, $p_b \in \mathbb{R}^d$ let

$$y_b^M = (\underbrace{0, \ldots, 0}_{M \text{ times}}, p_b, c_b) \in (\mathbb{R}^d)^M \times \mathbb{R}^d \times \mathbb{R}.$$

Interpretation: When $\Phi_b^M(L) = y_b^M$ for some $L \in U$, then L is consistent with the dynamical data and fulfils the normalisation conditions $\mathrm{Mm}(L)(\overline{x}_b) = p_b, L(\overline{x}_b) = c_b$. The condition $(\mathrm{EL}(L)(\overline{x}_b))_k = c_\tau$ of Proposition 1 is left out due to practical considerations that will be discussed later – see Remark 5.

4.1.3. Lagrangian as a conditional mean of Gaussian fields

Let us introduce the formulas required to infer a Lagrangian from data and predict uncertainty in the identified equations of motions and other linear observables such as Hamiltonian or symplectic structure. We postpone to Section 6 a more detailed derivation and a justification of applicability of the theory of Gaussian fields, such as the boundedness of certain operators. The following considers the noise-free case. We will make use of the following assumptions that are fulfilled when the observed system is governed by the Euler–Lagrange equations to a non-degenerate Lagrangian $L \in \mathcal{C}^2(\overline{\Omega})$ and when K is the square exponential kernel $K(\overline{x}, \overline{y}) = \exp(-\|x-y\|^2/l)$, $l > 0$ and Ω is a locally Lipschitz domain (Remark 8):

Assumption 1 *Assume that*

$$\{L \in \mathcal{C}^2(\overline{\Omega}) \mid \Phi_b^M(L) = y_b^M\} \cap U \neq \emptyset$$

and that the RKHS U to kernel K embeds continuously into $\mathcal{C}^2(\overline{\Omega})$. Let K be four times continuously differentiable.

By general theory recalled in Appendix A, the posterior distribution of the canonical Gaussian field ξ conditioned on the bounded[4] linear constraint $\Phi_b^M(L) = y_b^M$ is again a Gaussian field $\xi_M = \mathcal{N}(L, \mathcal{K}_{\Phi_b^M})$. It is characterised by the conditional mean L and the conditional covariance operator $\mathcal{K}_{\Phi_b^M}$. To compute L and $\mathcal{K}_{\Phi_b^M}$, define the symmetric matrix

$$\Theta \in \mathbb{R}^{((M+1)d+1)\times((M+1)d+1)}, \quad \Theta_{k,l} = (\Phi_b^M)_k \mathcal{K} (\Phi_b^M)_l, \quad 1 \leq k, l \leq (M+1)d+1,$$

where $(\Phi_b^M)_k$, $(\Phi_b^M)_l$ refer to the kth or lth component of Φ_b^M, respectively. In block matrix form, Θ can be written as

$$\Theta = \begin{pmatrix} (\mathrm{EL}^1_{\hat{x}^{(j)}} \mathrm{EL}^2_{\hat{x}^{(i)}} K)_{ij} & (\mathrm{EL}^1_{\hat{x}^{(j)}} \mathrm{Mm}^2_{\overline{x}_b} K)_j & (\mathrm{EL}^1_{\hat{x}^{(j)}} \mathrm{ev}^2_{\overline{x}_b} K)_j \\ (\mathrm{Mm}^1_{\overline{x}_b} \mathrm{EL}^2_{\hat{x}^{(i)}} K)_i & \mathrm{Mm}^1_{\overline{x}_b} \mathrm{Mm}^2_{\overline{x}_b} K & \mathrm{Mm}^1_{\overline{x}_b} \mathrm{ev}^2_{\overline{x}_b} K \\ (\mathrm{ev}^1_{\overline{x}_b} \mathrm{EL}^2_{\hat{x}^{(i)}} K)_i & \mathrm{ev}^1_{\overline{x}_b} \mathrm{Mm}^2_{\overline{x}_b} K & K(\overline{x}_b, \overline{x}_b). \end{pmatrix} \tag{18}$$

[4] $\Phi_b^M : C^2(\overline{\Omega}) \to \mathbb{R}^{(M+1)d+1}$ is bounded (Section 6.1.2).

The upper indices 1, 2 of the operator indicate their action on the first or second component of the kernel $K\colon \Omega \times \Omega \to \mathbb{R}$, i.e.

$$\mathrm{EL}^1_{\hat{x}(j)}\mathrm{EL}^2_{\hat{x}(i)}K = \mathrm{EL}_{\hat{x}(j)}\left(\overline{x} \mapsto \mathrm{EL}_{\hat{x}(i)}(\overline{y} \mapsto K(\overline{x},\overline{y}))\right) \quad \in \mathbb{R}$$

with analogous conventions for Mm and ev. Furthermore, we use the convention that when an operator EL, Mm, or ev is applied to functions with several components their application are understood component-wise. With

$$\mathcal{K}\Phi^M_b(\overline{x}) = \left(\mathrm{EL}_{\hat{x}(1)}K(\cdot,\overline{x}), \quad \dots \quad \mathrm{EL}_{\hat{x}(M)}K(\cdot,\overline{x}), \quad \mathrm{Mm}_{\overline{x}_b}K(\cdot,\overline{x}), \quad K(\overline{x}_b,\overline{x})\right)^\top$$

the conditional mean L of the posterior process ξ_M is given as

$$L = {y^M_b}^\top \Theta^\dagger \mathcal{K}\Phi^M_b, \tag{19}$$

where $\Theta^\dagger$ denotes the pseudo-inverse of Θ. The conditional covariance operator $\mathcal{K}_{\Phi^M_b}\colon U^* \to U$ is given by

$$\psi\mathcal{K}_{\Phi^M_b}\phi = \psi\mathcal{K}\phi - (\psi\mathcal{K}{\Phi^M_b}^\top)\Theta^\dagger(\Phi^M_b\mathcal{K}\phi) \tag{20}$$

for any $\psi, \phi \in U^*$. Here

$$\begin{aligned}
\psi\mathcal{K}_{\Phi^M_b}\phi &= \psi^1\phi^2K\\
\psi\mathcal{K}{\Phi^M_b}^\top &= \left(\psi^1\mathrm{EL}^2_{\hat{x}(2)}K, \quad \dots \quad \psi^1\mathrm{EL}^2_{\hat{x}(n)}K, \quad \psi^1\mathrm{Mm}^2_{\overline{x}_b}K, \quad \psi^1K(\cdot,\overline{x}_b)\right)\\
\Phi^M_b\mathcal{K}\phi &= \left(\mathrm{EL}^1_{\hat{x}(2)}\phi^2K, \quad \dots \quad \mathrm{EL}^1_{\hat{x}(n)}\phi^2K, \quad \mathrm{Mm}^1_{\overline{x}_b}\phi^2K \quad \phi^2K(\overline{x}_b,\cdot)\right)^\top.
\end{aligned}$$

Again, the upper indices 1, 2 of the linear functionals $\phi, \psi \in U^*$ denote actions on the first or second component of K, respectively.

The expressions ${y^M_b}^\top\Theta^\dagger$ and $\Theta^\dagger(\Phi^M_b\mathcal{K}\phi)$ in (19) and (20), respectively, are least-square solutions to the linear systems

$$\Theta z = y^M_b \quad \text{and} \quad \Theta Z = \Phi^M_b\mathcal{K}\phi \tag{21}$$

for z and Z. It is argued in Appendix A.2 and Appendix A.3 that these systems are solvable and that (19) and (25) are valid. Moreover, $\Theta^\dagger(\Phi^M_b\mathcal{K}\phi)$ and $\Theta^\dagger(\Phi^M_b\mathcal{K}\phi)$ can be substituted by any solution to the linear systems above without changing L in (19) or $\psi\mathcal{K}_{\Phi^M_b}\phi$ in (20).

Remark 3 (Computational aspects) The size of the linear systems (21) scales linearly with the number of data points and the dimension of the state-space. Thus the numerical complexity of solving the linear systems scales approximately cubically, when a direct method is used. The growth in computational complexity is typical for Gaussian process or kernel-based methods [49]. To tackle this, various approaches exist such as using kernels of finite band-width to promote sparsity of Θ, importance sampling, and sparse Gaussian processes which are based on identifying inducing variables [55, 47]. An efficient method to approximate Cholesky factors of covariance matrices was presented in [53]. Moreover, a diagonal regularisation technique involving an adaptive nudging term can be found in [12, Appendix A] in the context of solving pdes with Gaussian processes. A more specialised approach is [54]. In our numerical experiments (Section 5) we do not employ any specialised algorithm but use the command `factorize` of the package Julia/LinearAlgebra [8] on Θ. Depending on the degeneracy of the symmetric matrix Θ, `factorize` computes a Cholesky decomposition or a factorisation based on the Bunch-Kaufman algorithm [5, 9]. The factors are then stored and used whenever solving linear systems involving Θ. □

Remark 4 (Equivalent minimisation problem) The conditional mean L of (19) can alternatively be characterised as the minimiser of the following convex optimisation problem

$$L = \arg \min_{\tilde{L} \in U, \Phi_b^M(\tilde{L}) = y_b^M} \|\tilde{L}\|_U, \tag{22}$$

where $\|\tilde{L}\|_U$ denotes the reproducing kernel Hilbert space norm. (See Theorem 8 in Appendix A.) This will play an important role in the convergence proof in Section 6. Besides the exploit for convergence proofs, formulation (22) could be used for the computation of the conditional stochastic processes for non-linear observations and normalisation conditions such as in the alternative regularisation of Appendix B using techniques of [12]. □

Remark 5 (Further normalisation) For consistency with Proposition 1, one may add $c_\tau \in \mathbb{R} \setminus \{0\}$ to y_b^M and the normalising condition $(\mathrm{EL}_{(\hat{x}_\tau)})_k$ to Φ_b^M for $\hat{x}_\tau = (x_\tau, \dot{x}_\tau, \ddot{x}_\tau)$ that is not a motion and $k \in \{1, \ldots, d\}$. While it is realistic to assume knowledge of a data point $\hat{x}_\tau$ that is not a motion (e.g. $\hat{x} = (\overline{x}^{(1)}, \ddot{x}^{(1)} + 1)$ in systems with non-degenerate true Lagrangian), fixing an index k a priori may cause a restriction as to which Lagrangians can be approximated or cause poor scaling of the posterior process. Thus, we propose to leave out this condition in the definition of the posterior process. One may rather verify $c_\tau \neq 0$ a posteriori to check validity of the assumptions of Proposition 1. Moreover, Appendix B discusses an alternative normalisation based on symplectic volume forms. It can be compared to approaches to learn Lagrangians with neural networks [42]. □

4.1.4. Application

The conditional mean L (19) of the posterior Gaussian process $\xi|_{\Phi_b^M(L)=y_b^M}$ serves as an approximation to a true Lagrangian, from which approximations of geometric structures such as symplectic structure and Hamiltonians can be derived. Moreover, uncertainties of a linear observables $\psi \in U^*$ can be quantified as the variance of $\psi(\xi|_{\Phi_b^M(L)=y_b^M})$, which can be computed as $\psi \mathcal{K}_{\Phi_b^M} \psi$ using (20). In the numerical experiments, standard deviations will be computed for the random variables $\mathrm{Ham}(\xi|_{\Phi_b^M(L)=y_b^M})(\overline{x})$ for $\overline{x} \in \Omega$ and for $\mathrm{EL}(\xi|_{\Phi_b^M(L)=y_b^M})(\hat{x}(t))$, where $\hat{x} = (x, \dot{x}, \ddot{x})$ is a motion of the approximate system to L.

4.2. Gaussian fields for discrete Lagrangians

The data-driven framework for learning of discrete Lagrangians is in close analogy to the presented framework for continuous Lagrangians. Instead of repeating the discussion, we explain the required modifications and reinterpretations in the following. A rigorous discussion and justification of the applicability of the theory of Gaussian fields is postponed to Section 6.2.

In the setting of discrete Lagrangians, $\Omega \subset \mathbb{R}^d \times \mathbb{R}^d$ is an open, bounded subset containing elements denoted by $\overline{x} = (x_0, x_1)$. Observed data corresponds to a collection of M triples of snapshots $\hat{x}^{(j)} = (x_0^{(j)}, x_1^{(j)}, x_2^{(j)})$ of motions of a variational dynamical system, where $(x_0^{(j)}, x_1^{(j)}) \in \Omega$ and $(x_1^{(j)}, x_2^{(j)}) \in \Omega$ for all j. The snapshot time (discretisation parameter) $\Delta_t > 0$ is constant (also see Figure 7). The goal is to identify a discrete Lagrangian $L_d \colon \Omega \to \mathbb{R}$ such that discrete motions that fulfil the discrete Euler-Lagrange equations $\mathrm{DEL}(L_d) = 0$ approximate true motions. When a system is governed by a non-degenerate continuous Lagrangian L, then there exists a discrete Lagrangian L_d that exactly governs the discretised dynamics [35].

Consider a twice countinuously differentiable kernel $K \colon \Omega \times \Omega \to \mathbb{R}$ with RKHS U. We consider the following assumptions that are fulfilled when the observed system is governed by the Euler–Lagrange equations to a non-degenerate Lagrangian $L \in \mathcal{C}^1(\overline{\Omega})$ and when K is the square exponential kernel $K(\overline{x}, \overline{y}) = \exp(-\|x - y\|^2/l)$, $l > 0$ and Ω is a locally Lipschitz domain:

Assumption 2 *Assume that*

$$\{L_d \in \mathcal{C}^1(\overline{\Omega}) \,|\, \Phi_b^M(L) = y_b^M\} \cap U \neq \emptyset$$

and that the RKHS U to kernel K embeds continuously into $\mathcal{C}^1(\overline{\Omega})$. Let K be twice continuously differentiable.

With the reinterpretation of Ω and of training data points $\hat{x}^{(j)}$ we can follow the framework for continuous Lagrangians replacing EL by DEL and Mm by Mm^- (or Mm^+). In particular, this leads to

$$\Phi_b^M = (\mathrm{DEL}_{\hat{x}^{(1)}}, \ldots, \mathrm{DEL}_{\hat{x}^{(M)}}, \mathrm{Mm}^-{}_{\overline{x}_b}, \mathrm{ev}_{\overline{x}_b}).$$

(cf. (17)) and

$$\Theta = \begin{pmatrix} (\mathrm{DEL}^1_{\hat{x}^{(j)}} \mathrm{DEL}^2_{\hat{x}^{(i)}} K)_{ij} & (\mathrm{DEL}^1_{\hat{x}^{(j)}} \mathrm{Mm}^{-2}_{\overline{x}_b} K)_j & (\mathrm{DEL}^1_{\hat{x}^{(j)}} \mathrm{ev}^2_{\overline{x}_b} K)_j \\ (\mathrm{Mm}^{-1}_{\overline{x}_b} \mathrm{DEL}^2_{\hat{x}^{(i)}} K)_i & \mathrm{Mm}^{-1}_{\overline{x}_b} \mathrm{Mm}^{-2}_{\overline{x}_b} K & \mathrm{Mm}^{-1}_{\overline{x}_b} \mathrm{ev}^2_{\overline{x}_b} K \\ (\mathrm{ev}^1_{\overline{x}_b} \mathrm{DEL}^2_{\hat{x}^{(i)}} K)_i & \mathrm{ev}^1_{\overline{x}_b} \mathrm{Mm}^{-2}_{\overline{x}_b} K & K(\overline{x}_b, \overline{x}_b). \end{pmatrix} \tag{23}$$

(cf. (18)) and an a conditioned process that is a Gaussian process $\mathcal{N}(L, \mathcal{K}_{\Phi_b^M})$ with posterior mean

$$L_d = {y_b^M}^\top \Theta^\dagger \mathcal{K} \Phi_b^M \tag{24}$$

(cf. (19)). Again, the upper index 1, 2 of the operators DEL, Mm^-, ev denote on which input element of K they act. The conditional covariance operator $\mathcal{K}_{\Phi_b^M} : U^* \to U$ is defined for any $\psi, \phi \in U^*$ by

$$\psi \mathcal{K}_{\Phi_b^M} \phi = \psi \mathcal{K} \phi - (\psi \mathcal{K} {\Phi_b^M}^\top) \Theta^\dagger (\Phi_b^M \mathcal{K} \phi). \tag{25}$$

Here

$$\begin{aligned}
\psi \mathcal{K}_{\Phi_b^M} \phi &= \psi^1 \phi^2 K \\
\psi \mathcal{K} {\Phi_b^M}^\top &= \left(\psi^1 \mathrm{DEL}^2_{\hat{x}^{(2)}} K, \quad \ldots \quad \psi^1 \mathrm{DEL}^2_{\hat{x}^{(n)}} K, \quad \psi^1 \mathrm{Mm}^{-2}_{\overline{x}_b} K, \quad \psi^1 K(\cdot, \overline{x}) \right) \\
\Phi_b^M \mathcal{K} \phi &= \left(\mathrm{DEL}^1_{\hat{x}^{(2)}} \phi^2 K \quad \ldots \quad \mathrm{DEL}^1_{\hat{x}^{(n)}} \phi^2 K \quad \mathrm{Mm}^{-1}_{\overline{x}_b} \phi^2 K \quad \phi^2 K(\overline{x}, \cdot) \right)^\top .
\end{aligned}$$

To obtain (24) and (25) we have (as in the continuous case) applied general theory as recalled in Proposition 8 in Appendix A.2. Indeed, conditions for the applicability of Proposition 8 are verified in Proposition 10 (Appendix A.2).

5. Numerical experiments

5.1. Continuous Lagrangians

Consider dynamical data $\hat{x}^{(j)} = (x^{(j)}, \dot{x}^{(j)}, \ddot{x}^{(j)})$, $j = 1, \ldots, M$ of the coupled harmonic oscillator $L_{\mathrm{ref}} : T\mathbb{R}^2 \to \mathbb{R}$ with

$$L_{\mathrm{ref}}(x, \dot{x}) = \frac{1}{2}\|\dot{x}\|^2 - \frac{1}{2}\|x\|^2 + \alpha x^0 x^1, \quad x = (x^0, x^1) \in \mathbb{R}^2, (x, \dot{x}) \in T\mathbb{R}^2 \tag{26}$$

with coupling constant $\alpha = 0.1$. Here $\overline{x}^{(j)} = (x^{(j)}, \dot{x}^{(j)})$, $j = 1, \ldots, M$ are the first M elements of a Halton sequence in the hypercube $\Omega = [-1, 1]^4 \subset T\mathbb{R}^2$. We use radial basis functions $K(\overline{x}, \overline{y}) = \exp\left(-\frac{1}{2}(\overline{x} - \overline{y})^2\right)$ as a kernel function in all experiments. For $M \in \mathbb{N}$ we obtain a posteriori Gaussian processes denoted by $\xi_M \in \mathcal{N}(L_M, \mathcal{K}_M)$ modelling Lagrangians for the dynamical system. We present experiments with $M \in \{80, 300\}$. In the following var refers to the variance of a random variable (applied component wise when the random variable is $\mathbb{R}^d$-valued). Moreover, $\mathrm{Acc}_{\overline{x}}(L_M)$ refers to the solution of $\mathrm{EL}(L_M)(\overline{x}, \ddot{x}) = 0$ for $\ddot{x} \in \mathbb{R}^2$.

Figure 1 displays the location of training data in Ω projected to the (x^0, x^1)-plane. Figure 2 compares the variances of $\mathrm{EL}_{\hat{x}}(\xi_M)$ for $M = 80, 300$ for points of the form $\hat{x} = (\overline{x}, \ddot{x})$ with $\overline{x} = (x^0, x^1, 0, 0) \in \Omega$ and $\overline{x} = (x^0, 0, \dot{x}^0, 0) \in \Omega$ with $\ddot{x} = \mathrm{Acc}_{\overline{x}}(L_M)$. One observes that the variance decreases as more data points are used. This experiments suggests that the method can be used in combination with an adaptive sampling technique to sample new data points in regions of high model uncertainty. However, for consistency, our data points are related to a Halton sequence.

Figure 3 shows a motion computed by solving[5] $\mathrm{EL}(L_M) = 0$ with initial data $\overline{x} = (0.2, 0.1, 0, 0)$ on the time interval $[0, 100]$. In the plots of the first row, colours indicate

[5] Computations were performed using DifferentialEquations.jl[48]. Comparison with a trajectory computed using the variational midpoint rule [35] (step-size $h = 0.01$) shows a maximal difference in the x-component smaller than 3.5×10^{-4} ($M = 300$) along the trajectory.

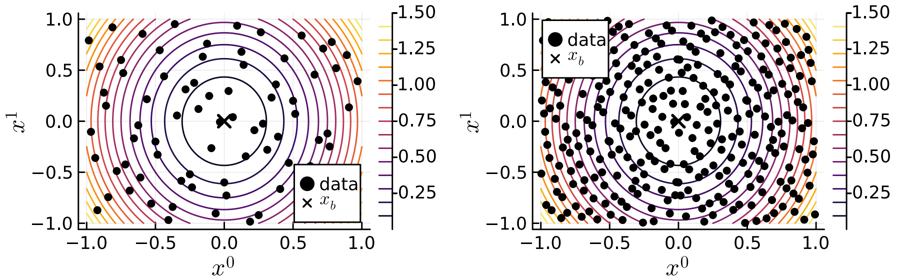

Figure 1: Training data points projected to the (x^0, x^1)-plane of ξ_{80} (left) and ξ_{300} (right).

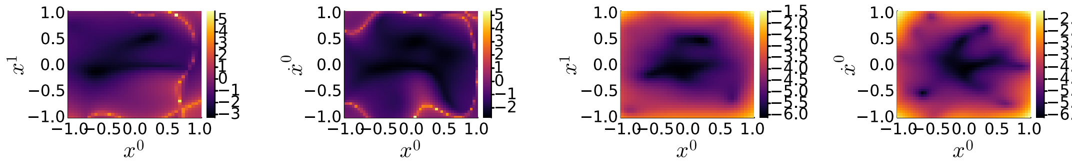

Figure 2: Plots of variances $\log_{10}(\|\mathrm{var}(\mathrm{EL}(\xi_M))\|)$ for $M = 80$ (two left plots) and $M = 300$ (two right plots) over $(x^0, x^1, 0, 0)$-plane and $(x^0, 0, \dot{x}^0, 0)$-plane. (Ranges of colourbars vary.)

the norm of the variance of $\mathrm{EL}(\xi_M)$ along the computed trajectories. For $M = 300$ the trajectory is close to the reference solution while largely different for $M = 80$. This is consistent with the lower variance for $M = 300$ compared to the experiment with $M = 80$. The plots of the dynamics of L_{300} (bottom row of Figure 3) show divergence of the computed motion from the reference solution towards the end of the time interval building up to a difference in x^0 component of about 0.1 at $t = 100$. (We will see later that a discrete model model performs better in this experiment.) However, the qualitative features of the motion are captured.

Figure 4 shows the Hamiltonian $H_M = \mathrm{Ham}(L_M)$ as well as $H_M \pm 0.2\sigma_{H_M}$. Here σ_{H_M} denotes the standard deviation $\sqrt{\mathrm{varHam}(\xi_M)}$. We observe a clear decrease of the standard deviation as M increases from 80 to 300.

Figure 5 displays the error in the prediction of $\ddot{x}$ for points $\overline{x} = (x^0, x^1, 0, 0) \in \Omega$ and $\overline{x} = (x^0, 0, \dot{x}^0, 0) \in \Omega$. As the magnitudes of errors vary widely, $\log_{10}$ is applied before plotting, i.e. we show the quantity

$$\log_{10} \|\mathrm{Acc}_{\overline{x}}(L_M) - \mathrm{Acc}_{\overline{x}}(L_{\mathrm{ref}})\|_{\mathbb{R}^2}.$$

One sees a clear decrease in error as M is increased from 80 to 300.

Figure 6 shows a convergence plot for the relative error in predicted acceleration $\mathrm{err}_{\mathrm{Acc}}$, i.e. of

$$\mathrm{err}_{\mathrm{Acc}}(\overline{x}) = \frac{\|\mathrm{Acc}_{\overline{x}}(L_M) - \mathrm{Acc}_{\overline{x}}(L_{\mathrm{ref}})\|_{\mathbb{R}^d}}{\|\mathrm{Acc}_{\overline{x}}(L_{\mathrm{ref}})\|_{\mathbb{R}^d}}.$$

The data for the plot in Figure 6 was computed for the 1d harmonic oscillator $L_{\mathrm{ref}}(x) = \frac{1}{2}\dot{x}^2 - \frac{1}{2}x^2$ with $(x, \dot{x}) \in [-1, 1]^2$ in quadruple precision. For each $M \in \{2^1, 2^2, \ldots, 2^6\}$

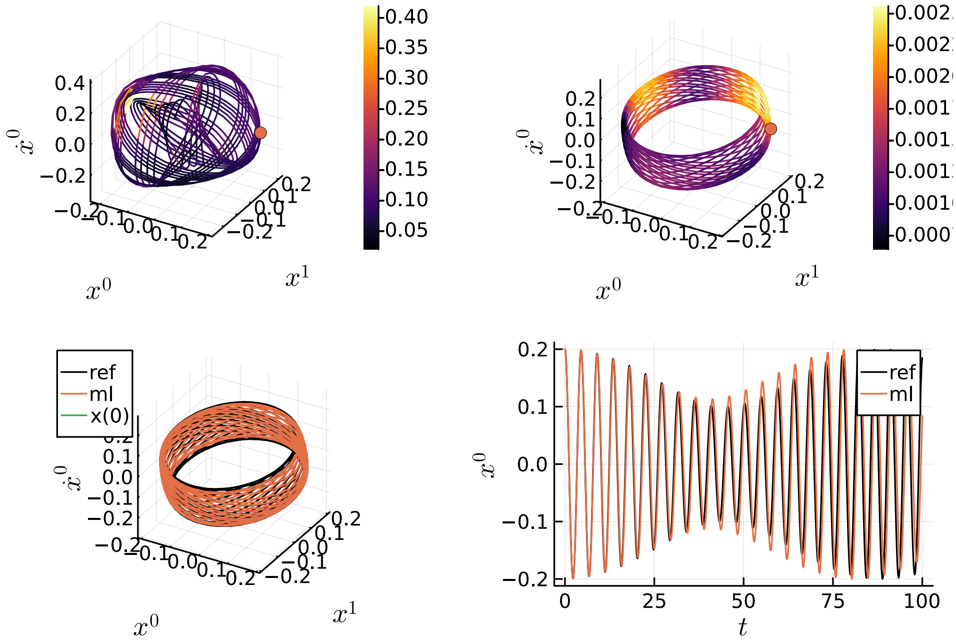

Figure 3: Top row: motion of ξ_{80} (left) and ξ_{300} (right) with variance $\|\mathrm{var}(\mathrm{EL}(\xi_M))\|$ encoded as colours (ranges of colourbars vary). Bottom row: motions of ξ_{300} compared to reference.

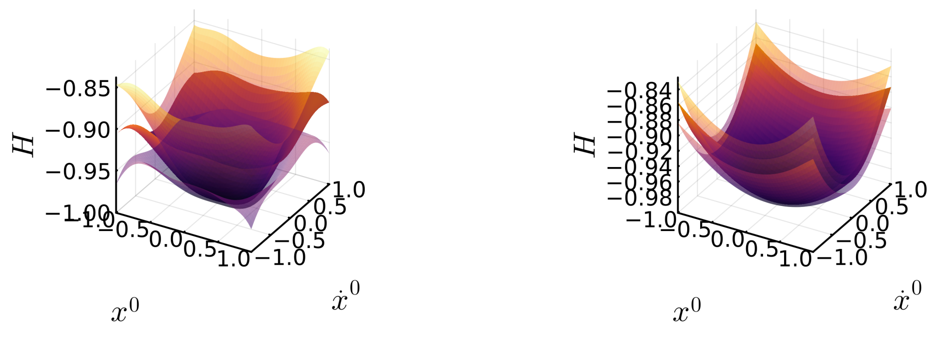

Figure 4: Mean of Hamiltonian $\mathrm{Ham}(\xi_{80})$, $\mathrm{Ham}(\xi_{300})$ over $(x^0, 0, \dot{x}^0, 0)$ plus/minus 20% standard deviation.

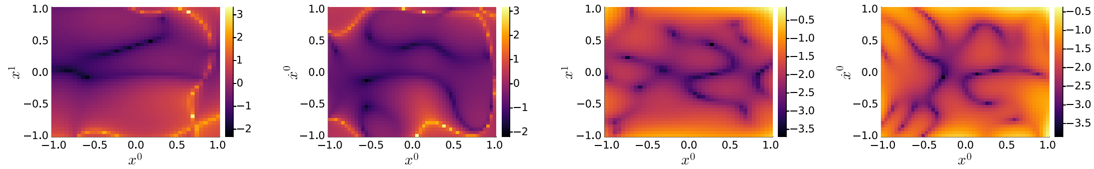

Figure 5: $\log_{10}$ norm of error of predicted acceleration $\ddot{x}$ for $\mathrm{Acc}(\xi_M)$ over x^0, x^1 plane and $x^0, \dot{x}^0$ plane for $M = 80$ (left two plots) and $M = 300$ (right two plots). (The ranges of colourbars vary.)

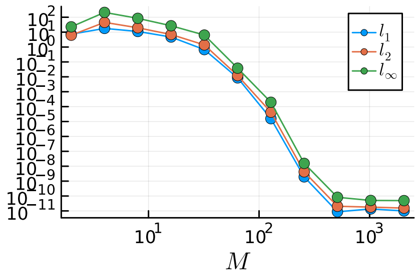

Figure 6: Convergence of $\mathrm{Acc}(L_M)$ to true acceleration data.

the error $\mathrm{err}_{\mathrm{Acc}}(\overline{x})$ was evaluated on a uniform mesh with 10×11 mesh points in $[-1, 1] \times [-1, 1] \in T\mathbb{R}$. The plot shows the discrete L_p error ($p = 1, 2, \infty$). We can see convergence with errors levelling out due to round-off errors at approximately 10^{-11}. Moreover, as M increases, higher and higher convergence rates become dominant before round-off errors dominate. Indeed, our analysis on convergence rates (Section 6.3) will show that thanks to smoothness of the kernel and the reference Lagrangian, arbitrarily high convergence rates occur in the asymptotic regime $M \to \infty$ (Corollary 3).

5.2. Discrete Lagrangian

Now we consider dynamical data $\hat{x}^{(j)} = (x_0^{(j)}, x_1^{(j)}, x_2^{(j)})$ where $x_0^{(j)}$, $x_1^{(j)}$, $x_2^{(j)}$ are snapshots of true trajectories at times t, $t+h$, $t+2h$, respectively, with $j = 1, \ldots, M$. Here $h = 0.1$ and, again, $M \in \{80, 300\}$. For data generation, we consider data $(x, p) \in [-1, 1]^4 \subset T^*\mathbb{R}^2$ from a Halton sequence from where we integrate L_{ref} from $[0, 3h]$ using the 2nd order accurate variational midpoint rule [35] with step-size $h_{\mathrm{internal}} = h/10$. These dynamics are considered as true for the purpose of this experiment. Training data is visualised in Figure 7.

Figure 8 (in analogy to Figure 2) shows how variance decreases as more data points become available. For the plots, $(x_0, p_0) \in T^*\mathbb{R}^2$ are used to compute $\hat{x} = (x_0, x_1, x_2)$ using L_{ref}. Here p refers to the conjugate momentum of L_{ref}. The plots display heatmaps of $\log_{10}(\|\mathrm{var}(\mathrm{DEL}_{\hat{x}}(\xi_M))\|)$.

Figure 9 shows a motion for $t \in [0, 100]$ of ξ_{300} with same initial data as in Figure 3.

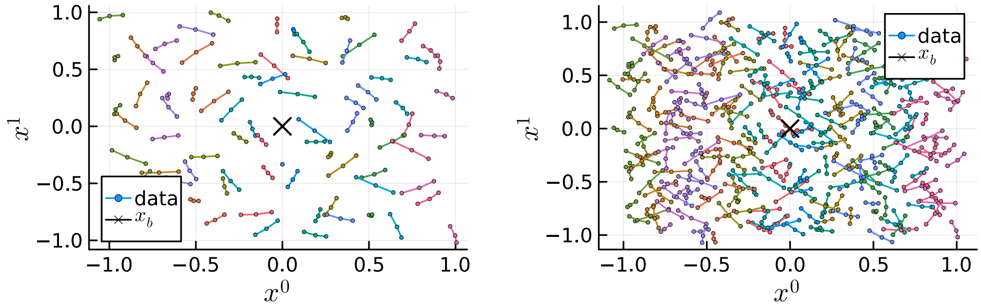

Figure 7: Training data. Each line connects snapshots points that constitute a training data point $\hat{x}$. Left: $M = 80$, right: $M = 300$.

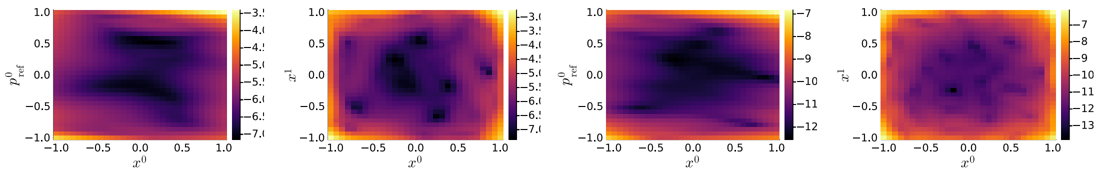

Figure 8: Plots of variances $\log_{10}(\|\mathrm{var}(\mathrm{EL}(\xi_M))\|)$ for $M = 80$ (left two plots) and $M = 300$ (right two plots) over $(x, p_{\mathrm{ref}}) = (x^0, x^1, 0, 0)$-plane and $(x, p_{\mathrm{ref}}) = (x^0, 0, p^0_{\mathrm{ref}}, 0)$-plane. (Ranges of colourbars vary.)

With a maximal error in absolute norm smaller than 0.00043 it is visually indistinguishable from the true motion. In the plot to the left, data for $\dot{x}^0$ was approximated to second order accuracy in h with the central finite differences method.

Comparing Figure 9 and Figure 3, it is interesting to observe that with the same amount of data the discrete model performs better than the continuous model for predicting motions.

Reproducibility Source code of the experiments can be found at `https://github.com/Christian-Offen/Lagrangian_GP`.

6. Convergence Analysis

This section contains a theoretical convergence analysis of the considered methods. In Sections 6.1 and 6.2 convergence theorems for regular continuous Lagrangians (Theorem 1) and discrete Lagrangians (Theorem 2) in the infinite-data limit are provided as observations become topologically dense, i.e. as the maximal distance between data points converges to zero. Moreover, the convergence rates of continuous and discrete Lagrangian models are analysed in Section 6.3.

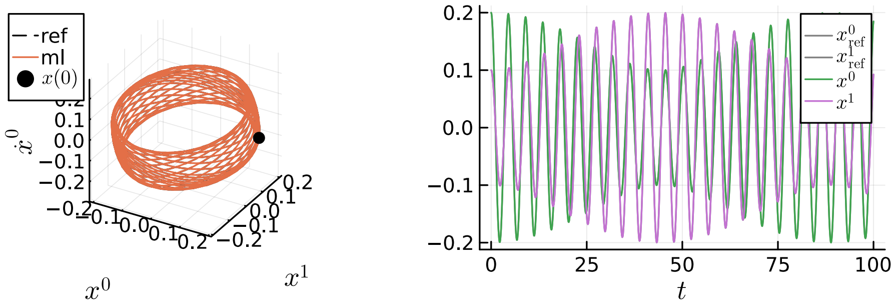

Figure 9: The motion of ξ_{300} and the true motion are indistinguishable.

6.1. Convergence of continuous Lagrangian models

6.1.1. Convergence theorem (continuous, temporal evolution)

Theorem 1 *Let $\Omega \subset T\mathbb{R}^d \cong \mathbb{R}^d \times \mathbb{R}^d$ be an open, bounded non-empty domain. Consider a sequence of observations $\Omega_0^E = \{(x^{(j)}, \dot{x}^{(j)}, \ddot{x}^{(j)})\}_j \subset \Omega \times \mathbb{R}^d$ of a dynamical system governed by the Euler–Lagrange equation of an (unknown) non-degenerate Lagrangian $L_{\mathrm{ref}} \in \mathcal{C}^2(\overline{\Omega})$ (definition of $\mathcal{C}^2(\overline{\Omega})$ below). Assume that $\{(x^{(j)}, \dot{x}^{(j)})\}_j \subset \Omega$ is topologically dense. Let K be a 4-times continuously differentiable kernel on Ω, $\overline{x_b} \in \Omega$, $r_b \in \mathbb{R}$, $p_b \in \mathbb{R}$ and assume that L_{ref} is contained in the reproducing kernel Hilbert space $(U, \|\cdot\|_U)$ to K and fulfils the normalisation condition*

$$\Phi_N(L_{\mathrm{ref}}) = (p_b, r_b) \quad \text{with} \quad \Phi_N(L) = \left(\frac{\partial L}{\partial \dot{x}}(\overline{x}_b), L(\overline{x}_b) \right). \tag{27}$$

Assume that U embeds continuously into $\mathcal{C}^2(\overline{\Omega})$. Let $\xi \in \mathcal{N}(0, \mathcal{K})$ be a canonical Gaussian process on U (see Section 4.1.1). Then the sequence of conditional means $L_{(j)}$ of ξ conditioned on the first j observations and the normalisation conditions

$$\mathrm{EL}(\xi)(x^{(i)}, \dot{x}^{(i)}, \ddot{x}^{(i)}) = 0 \, (\forall i \leq j), \quad \Phi_N(\xi) = (p_b, r_b) \tag{28}$$

converges in $\|\cdot\|_U$ and in $\|\cdot\|_{\mathcal{C}^2(\overline{\Omega})}$ to a Lagrangian $L_{(\infty)} \in U$ that is

- *consistent with the normalisation $\Phi_N(L_{(\infty)}) = (p_b, r_b)$*
- *consistent with the dynamics, i.e. $\mathrm{EL}(L_{(\infty)})(\hat{x})$ for all $\hat{x} = (x, \dot{x}, \ddot{x})$ with $(x, \dot{x}) \in \Omega$ and $\mathrm{EL}(L_{\mathrm{ref}})(\hat{x}) = 0$.*
- *Moreover, $L_{(\infty)}$ is the unique minimiser of $\|\cdot\|_U$ among all Lagrangians with these properties.* □

Remark 6 If $r_b = 0$ and $p_b = 0$, then the sequence $L_{(j)}$ is constantly zero with limit $L_{(\infty)} \equiv 0$. It is necessary to set $(r_b, p_b) \neq (0, 0)$ to approximate a non-degenerate Lagrangian. □

Remark 7 The regularity assumptions of the kernel (four times continuously differentiable) is required for the interpretation of $L_{(j)}$ as a conditional mean of a Gaussian process and for its convenient computation. It can be relaxed to the condition that $\frac{\partial^{|\alpha|+|\beta|}K(x,y)}{(\partial x^1)^{\alpha_1}\ldots(\partial x^d)^{\alpha_d}\partial(y^1)^{\beta_1}\ldots(\partial y^d)^{\beta_d}}$ for $|\alpha|,|\beta| \le 2$, $x,y \in \Omega$ exists and is continuous on $\overline{\Omega}$. Here $|\alpha| = \alpha_1 + \ldots + \alpha_d$, $|\beta| = \beta_1 + \ldots + \beta_d$. □

6.1.2. Formal setting and proof (continuous, temporal evolution)

Let $\Omega \subset T\mathbb{R}^d$ be an open, bounded, non-empty domain. Following notion of [1], we consider the space of m-times continuously differentiable functions that extend to the topological closure $\overline{\Omega}$

$$\mathcal{C}^m(\overline{\Omega},\mathbb{R}^k) = \{f \in \mathcal{C}^m(\Omega,\mathbb{R}^k) \mid \partial^\alpha f \text{ extends continuously to } \overline{\Omega}\ \forall |\alpha| \le m\}, \quad m \in \mathbb{N}_0.$$

Here $\partial^\alpha f = \frac{\partial^{|\alpha|} f}{(\partial x^1)^{\alpha_1}\ldots(\partial x^d)^{\alpha_d}\partial(\dot{x}^1)^{\dot{\alpha}_1}\ldots(\partial \dot{x}^d)^{\dot{\alpha}_d}}$ denotes the partial derivative with respect to coordinates $\overline{x} = (x,\dot{x}) = (x^1,\ldots,x^d,\dot{x}^1,\ldots,\dot{x}^d)$ for a multi-index $\alpha = (\alpha_1,\ldots,\alpha_d,\dot{\alpha}_1,\ldots,\dot{\alpha}_d)$ with $|\alpha| = \alpha_1 + \ldots + \alpha_d + \dot{\alpha}_1 + \ldots + \dot{\alpha}_d$. The space is equipped with the norm

$$\|f\|_{\mathcal{C}^m(\overline{\Omega},\mathbb{R}^k)} = \max_{0\le|\alpha|\le m} \sup_{\overline{x}\in\Omega} \|\partial^\alpha f(\overline{x})\|. \tag{29}$$

Here $\|\partial^\alpha f(\overline{x})\|$ denotes the Euclidean norm on $\mathbb{R}^k$ for $|\alpha| = 1$ or an induced operator norm for $|\alpha| > 1$. The space $\mathcal{C}^m(\overline{\Omega},\mathbb{R}^k)$ is a Banach space [1, § 4]. We will use the shorthand $\mathcal{C}^m(\overline{\Omega}) = \mathcal{C}^m(\overline{\Omega},\mathbb{R}^1)$.

Assume that on a dense, countable subset $\Omega_0 = \{\overline{x}^{(j)} = (x^{(j)},\dot{x}^{(j)})\}_{j=1}^\infty \subset \Omega$ we have observations of acceleration data $\ddot{x}^{(j)}$ of a dynamical system generated by an (a priori unknown) Lagrangian $L_{\text{ref}} \in \mathcal{C}^2(\overline{\Omega})$, which is non-degenerate, i.e. for all $(x,\dot{x}) \in \overline{\Omega}$ the matrix $\frac{\partial^2 L_{\text{ref}}}{\partial\dot{x}\partial\dot{x}}(x,\dot{x})$ is invertible, and the induced function $g_{\text{ref}} \in \mathcal{C}^0(\overline{\Omega},\mathbb{R}^d)$ with

$$g_{\text{ref}}(x,\dot{x}) = \left(\frac{\partial^2 L_{\text{ref}}}{\partial\dot{x}\partial\dot{x}}(x,\dot{x})\right)^{-1}\left(\frac{\partial L_{\text{ref}}}{\partial x}(x,\dot{x}) - \frac{\partial^2 L_{\text{ref}}}{\partial x\partial\dot{x}}(x,\dot{x})\cdot\dot{x}\right) \tag{30}$$

recovers $\ddot{x}^{(j)} = g_{\text{ref}}(\overline{x}^{(j)}) = g_{\text{ref}}(x^{(j)},\dot{x}^{(j)})$.

Lemma 3 *The linear functional* $\Phi^{(\infty)}\colon \mathcal{C}^2(\overline{\Omega}) \to \mathcal{C}^0(\overline{\Omega},\mathbb{R}^d)$ *with*

$$\begin{aligned}\Phi^{(\infty)}(L)(x,\dot{x}) &= \mathrm{EL}(L)(x,\dot{x},g_{\text{ref}}(x,\dot{x}))\\ &= \frac{\partial^2 L}{\partial\dot{x}\partial\dot{x}}(x,\dot{x})\cdot g_{\text{ref}}(x,\dot{x}) + \frac{\partial^2 L}{\partial x\partial\dot{x}}(x,\dot{x})\cdot\dot{x} - \frac{\partial L}{\partial x}(x,\dot{x})\end{aligned} \tag{31}$$

is bounded. □

Proof A direct application of the triangle inequality shows

$$\|\Phi^{(\infty)}(L)\|_{\mathcal{C}^0(\overline{\Omega},\mathbb{R}^d)} \le \left(\|g_{\text{ref}}\|_{\mathcal{C}^0(\overline{\Omega},\mathbb{R}^d)} + \sup_{(x,\dot{x})\in\Omega}\|\dot{x}\| + 1\right)\|L\|_{\mathcal{C}^2(\overline{\Omega})}.$$

■

Since for each $\overline{x}$ the evaluation functional $\mathrm{ev}_{\overline{x}}\colon f \mapsto f(\overline{x})$ on $\mathcal{C}^0(\overline{\Omega}, \mathbb{R}^d)$ is bounded, the following functions constitute bounded linear functionals for $j \in \mathbb{N}$:

$$\Phi_j \colon \mathcal{C}^2(\overline{\Omega}) \to \mathbb{R}^d, \qquad \Phi_j(L) = \Phi^{(\infty)}(L)(\overline{x}^{(j)})$$
$$\Phi^{(j)} \colon \mathcal{C}^2(\overline{\Omega}) \to (\mathbb{R}^d)^j, \quad \Phi^{(j)} = (\Phi_1, \dots, \Phi_j).$$

For a reference point $\overline{x}_b \in \Omega$ and for $p_b \in \mathbb{R}^d$, $r_b \in \mathbb{R}$ we define the bounded linear functional

$$\Phi_N \colon \mathcal{C}^2(\overline{\Omega}) \to \mathbb{R}^{d+1}, \qquad \Phi_N(L) = \left(\frac{\partial L}{\partial \dot{x}}(\overline{x}_b), L(\overline{x}_b) \right), \tag{32}$$

related to our normalisation condition, the shorthands $\Phi_b^{(k)} = (\Phi_1, \dots, \Phi_k, \Phi_N)$ and $\Phi_b^{(\infty)} = (\Phi^{(\infty)}, \Phi_N)$, and the data

$$y^{(k)} = (0, \dots, 0, p_b, r_b) \in (\mathbb{R}^d)^k \times \mathbb{R}^d \times \mathbb{R}$$
$$y^{(\infty)} = (0, p_b, r_b) \in \mathcal{C}^0(\overline{\Omega}, \mathbb{R}^d) \times \mathbb{R}^d \times \mathbb{R}.$$

Assumption 3 *Assume that there is a Hilbert space U with continuous embedding $U \hookrightarrow \mathcal{C}^2(\overline{\Omega})$ such that*

$$\{L \in \mathcal{C}^2(\overline{\Omega}) \,|\, \Phi_b^{(\infty)}(L) = y^{(\infty)}\} \cap U \neq \emptyset.$$

In other words, U is assumed to contain a Lagrangian consistent with the normalisation and underlying dynamics.

The affine linear subspaces

$$A^{(j)} = \{L \in U \,|\, \Phi_b^{(j)}(L) = y^{(j)}\} \quad (j \in \mathbb{N})$$
$$A^{(\infty)} = \{L \in U \,|\, \Phi_b^{(\infty)}(L) = y^{(\infty)}\}$$

are closed and non empty in U by Assumption 3 and by the boundedness of $\Phi_b^{(j)}$ and $\Phi_b^{(\infty)}$ on $U \hookrightarrow \mathcal{C}^2(\overline{\Omega})$. Therefore, the following minimisation constitute convex optimisation problems on B with unique minima in $A^{(j)}$ or $A^{(\infty)}$, respectively:

$$\begin{aligned} L_{(j)} &= \arg \min_{L \in A^{(j)}} \|L\|_U \\ L_{(\infty)} &= \arg \min_{L \in A^{(\infty)}} \|L\|_U. \end{aligned} \tag{33}$$

Here $\| \cdot \|_U$ denotes the norm in U.

Remark 8 To an open, non-empty set $X \subset \mathbb{R}^d$, $m \in \mathbb{N} \cup \{0\}$ denote by $W^{m,2}(X) = W^m(X)$ the Sobolev space

$$W^m(X) = \{u \in L^2(X) \,|\, \forall \alpha \in \mathbb{N}^d, |\alpha| \leq m, \partial^\alpha u \in L^2(X)\},$$

with Sobolev norm

$$\|u\|_{W^m} = \sqrt{\sum_{|\alpha|\leq m} \int_X (\partial^\alpha u(x))^2 \mathrm{d}x}$$

where $L^2(X)$ denotes the space of square integrable functions on X. Here the derivative $\partial^\alpha u$ is meant in a distributional sense [1]. In the machine learning setting, U is the reproducing kernel Hilbert space related to a kernel $K\colon \Omega\times\Omega \to \mathbb{R}$. Assume the domain of Ω is locally Lipschitz. When K is the squared exponential kernel, for instance, its reproducing kernel Hilbert space embeds into any Sobolev space $W^m(\Omega)$ ($m > 1$) [14, Thm.4.48]. In particular , it embeds into $W^m(\Omega)$ with $m > 2 + d/2$, which is embedded into $\mathcal{C}^2(\overline{\Omega})$ by the Sobolev embedding theorem [1, §4]. The element L_j from (33) coincides with the conditional mean of the Gaussian process ξ conditioned on $\Phi_b^{(j)}(\xi) = y^{(j)}$ (Remark 4). $\square$

Proposition 5 *The minima $L_{(j)}$ converge to $L_{(\infty)}$ in the norm $\|\cdot\|_U$ and, thus, in $\|\cdot\|_{\mathcal{C}^2(\overline{\Omega})}$.* $\square$

Proof The sequence of affine spaces $A^{(1)} \supseteq A^{(2)} \supseteq A^{(3)} \supseteq \ldots$ is monotonously decreasing and $A^{(\infty)} \subseteq \bigcap_{j=1}^\infty A^{(j)}$. Therefore, the sequence $L_{(j)}$ is monotonously increasing and its norm $\|L_{(j)}\|_U$ is bounded from above by $\|L_{(\infty)}\|_U$. Since U is reflexive, there exists a subsequence $(L_{(j_i)})_{i\in\mathbb{N}}$ that weakly converges to some $L^\dagger_{(\infty)} \in U$. (This follows from the Banach-Alaoglu theorem and the Eberlein-Šmulian theorem [19].) By the weak lower semi-continuity of the norm, we obtain

$$\|L^\dagger_{(\infty)}\|_U \leq \liminf_{i\to\infty} \|L_{(j_i)}\|_U \leq \|L_{(\infty)}\|_U. \tag{34}$$

Lemma 4 *The weak limit $L^\dagger_{(\infty)}$ of $(L_{(j_i)})_{i\in\mathbb{N}}$ is contained in $A^{(\infty)}$.* $\square$

Before providing the proof of Lemma 4, we show how this allows us to complete the proof of Proposition 5.

As $L^\dagger_{(\infty)} \in A^{(\infty)}$, we have $\|L_{(\infty)}\|_U \leq \|L^\dagger_{(\infty)}\|_U$ since $L_{(\infty)}$ is the global minimiser of the minimisation problem of (33). Together with (34) we conclude $\|L^\dagger_{(\infty)}\|_U = \|L_{(\infty)}\|_U$ and, by the uniqueness of the minimiser $L_{(\infty)}$, the equality $L^\dagger_{(\infty)} = L_{(\infty)}$. Thus, we have proved weak convergence $L_{(j_i)} \rightharpoonup L_{(\infty)}$.

Together with the lower semi-continuity of the norm, and since $L_{(j_i)}$ is monotonously increasing and bounded by $\|L_{(\infty)}\|_U$, we have

$$\|L_{(\infty)}\|_U \leq \liminf_{i\to\infty} \|L_{(j_i)}\|_U \leq \limsup_{i\to\infty} \|L_{(j_i)}\|_U \leq \|L_{(\infty)}\|_U$$

such that $\lim_{i\to\infty} \|L_{(j_i)}\|_U = \|L_{(\infty)}\|_U$. Together with $L_{(j_i)} \rightharpoonup L_{(\infty)}$ we conclude strong convergence $L_{(j_i)} \to L_{(\infty)}$ in the Hilbert space U.

The particular weakly convergent subsequence $(L_{(j_i)})_{i\in\mathbb{N}}$ of $(L_{(j)})_j$ was arbitrary. Thus, any weakly convergent subsequence of $(L_{(j)})_j$ converges strongly against $L_{(\infty)}$.

It follows that any subsequence of $(L_{(j)})_j$ has a subsequence that converges to $L_{(\infty)}$. This implies that the whole series $(L_{(j)})_j$ converges to $L_{(\infty)}$.

It remains to prove Lemma 4.

PROOF (LEMMA 4) Let $\overline{x} \in \Omega$. As the sequence $\Omega_0 = (\overline{x}^{(m)})_{m=1}^{\infty}$ is dense in Ω, there exists a subsequence $(\overline{x}^{(m_l)})_{l=1}^{\infty}$ converging to $\overline{x}$. We have

$$\Phi_b^{(\infty)}(L_{(\infty)}^{\dagger})(\overline{x}) = \lim_{l\to\infty} \Phi_b^{(\infty)}(L_{(\infty)}^{\dagger})(\overline{x}^{(m_l)}) \tag{35}$$

$$= \lim_{l\to\infty} \underbrace{\lim_{i\to\infty} \Phi_b^{(\infty)}(L_{(j_i)})(\overline{x}^{(m_l)})}_{\overset{(*)}{=}0} = 0. \tag{36}$$

For this, in (35) we use that $\Phi_b^{(\infty)}(L_{(\infty)}^{\dagger}) \in \mathcal{C}^0(\overline{\Omega})$. Equality in (36) follows because each projection to a component of $\Phi_b^{(\infty)}(\cdot)(\overline{x}^{(m_l)}) \colon U \to \mathbb{R}^d \times \mathbb{R}^{d+1}$ constitutes a bounded linear functional on U and the sequence $(L_{(j_i)})_{i\in\mathbb{N}}$ converges weakly to $L_{(\infty)}^{\dagger}$. Finally, equality $(*)$ holds because for each l there exists $N \in \mathbb{N}$ such that $j_N \geq m_l$ and then for all $i \geq N$ we have $\Phi_b^{(\infty)}(L_{(j_i)})(\overline{x}^{(m_l)}) = 0$.

From $\Phi_b^{(\infty)}(L_{(\infty)}^{\dagger})(\overline{x}) = 0$ for all $\overline{x} \in \Omega$ we conclude $L_{(\infty)}^{\dagger} \in A^{(\infty)}$. ■

This completes the proof of Proposition 5. ■

Now we can easily prove Theorem 1:

PROOF (THEOREM 1) By Theorem 8 of Appendix A.2 (also see Remark 4) the conditional means computed in (19) coincide with the unique minimisers of the problems (33). Indeed, the assumption of Theorem 8 on $y = y_b^M$ is verified in Proposition 9 of Appendix A.3. Theorem 1 is, therefore, a direct consequence of Proposition 5. ■

6.2. Convergence of discrete Lagrangian models

6.2.1. Statement of convergence theorem (discrete, temporal evolution)

Theorem 2 *Let $\Omega_a, \Omega_b \subset \mathbb{R}^d \times \mathbb{R}^d$ be open, bounded, non-empty domains. Let $\Omega = \Omega_a \cup \Omega_b$. Consider a sequence of observations*

$$\hat{\Omega}_0 = \{\hat{x}^{(j)} = ({x_0}^{(j)}, {x_1}^{(j)}, {x_2}^{(j)})\}_{j=1}^{\infty}$$

of a discrete dynamical system with (not explicitly known) globally Lipschitz continuous discrete flow map $g \colon \Omega_a \to \Omega_b$ related to a discrete Lagrangian $L_d^{\mathrm{ref}} \in \mathcal{C}^1(\overline{\Omega})$, i.e.

- $g(x_0^{(j)}, x_1^{(j)}) = (x_1^{(j)}, x_2^{(j)})$ *for all* $j \in \mathbb{N}$,
- $\mathrm{DEL}(L_d^{\mathrm{ref}})(x_0, g(x_0, x_1)) = 0$ *for all* $(x_0, x_1) \in \Omega_a$,
- $\nabla_{1,2} L_d^{\mathrm{ref}}(x_1, x_2) \in \mathbb{R}^{d\times d}$ *is invertible for all* $(x_1, x_2) \in \overline{\Omega}_b$.

Assume that $\{(x_0{}^{(j)}, x_1{}^{(j)})\}_{j=1}^{\infty}$ *is dense in* Ω_a*. Let* K *be a twice continuously differentiable kernel on* Ω*,* $\sqcap_b \in \Omega$*,* $r_b \in \mathbb{R}$*,* $p_b \in \mathbb{R}$ *and assume that* L_d^{ref} *is contained in the reproducing kernel Hilbert space* $(U, \|\cdot\|_U)$ *to* K *and fulfils the normalisation condition*

$$\Phi_N(L_d^{\mathrm{ref}}) = (p_b, r_b) \quad \textit{with} \quad \Phi_N(L_d) = (-\nabla_2 L_d(\sqcap_b), L_d(\sqcap_b)) \tag{37}$$

and that U *embeds continuously into* $\mathcal{C}^1(\overline{\Omega})$*. Let* $\xi \in \mathcal{N}(0, \mathcal{K})$ *be a centred Gaussian random variable over* U*. Then the sequence of conditional means* $L_{d,(j)}$ *of* ξ *conditioned on the first* j *observations and the normalisation conditions*

$$\mathrm{DEL}(\xi)(\hat{x}^{(i)}) = 0 \, (\forall i \leq j), \quad \Phi_N(\xi) = (p_b, r_b) \tag{38}$$

converges in $\|\cdot\|_U$ *and in* $\|\cdot\|_{\mathcal{C}^1(\overline{\Omega})}$ *to a Lagrangian* $L_{d,(\infty)} \in U$ *that is*

- *consistent with the normalisation* $\Phi_N(L_{d,(\infty)}) = (p_b, r_b)$
- *consistent with the dynamics, i.e.* $\mathrm{DEL}(L_{d,(\infty)})(\hat{x}) = 0$ *for all* $\hat{x} = (x_0, x_1, x_2)$ *with* $(x_0, x_1) \in \Omega_a$*,*$(x_1, x_2) \in \Omega_b$ *and* $\mathrm{DEL}(L_d^{\mathrm{ref}})(\hat{x}) = 0$*.*
- *Moreover,* L_d *is the unique minimizer of* $\|\cdot\|_U$ *among all discrete Lagrangians in* U *with the properties above.* □

Remark 9 The regularity assumption of K (twice continuously differentiable) is needed for the interpretation of $L_{d,(j)}$ as a conditional mean of a Gaussian process and for a convenient computation of $L_{d,(j)}$. However, the proof will show that a relaxation to continuous differentiability is possible. □

6.2.2. Formal setting and proof (discrete, temporal evolution)

Let $\Omega_a, \Omega_b \subset \mathbb{R}^d \times \mathbb{R}^d$ be open, bounded, non-empty domains, let $\Omega = \Omega_a \cup \Omega_b$. Let $\hat{\Omega} = \{(x_0, x_1, x_2) \,|\, (x_0, x_1) \in \Omega_a, (x_1, x_2) \in \Omega_b\}$ and let

$$\hat{\Omega}_0 = \{(x_0^{(j)}, x_1^{(j)}, x_2^{(j)})\}_{j=1}^{\infty} \subset \hat{\Omega} \quad \text{with } (x_0^{(j)}, x_1^{(j)}) \in \Omega_a, (x_1^{(j)}, x_2^{(j)}) \in \Omega_b \text{ for all } j \in \mathbb{N}.$$

Assume that $\{(x_0^{(j)}, x_1^{(j)})\}_{j=1}^{\infty}$ is dense in Ω_a.

Remark 10 (Interpretation of $\hat{\Omega}_0$) The set $\hat{\Omega}_0$ corresponds to a collection of observation data in the infinite data limit. It can be obtained as a collection of three consecutive snapshots of motions of the dynamical system that we observe and for which we seek to learn a discrete Lagrangian. In a typical scenario where $L_d^{\mathrm{ref}} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the exact discrete Lagrangian to some underlying continuous Lagrangian, the motions leave the diagonal of $\mathbb{R}^d \times \mathbb{R}^d$ invariant. It is sensible to consider Ω_a and Ω_b that are neighbourhoods of compact sections of the diagonal in $\mathbb{R}^d \times \mathbb{R}^d$. □

We consider the discrete Lagrangian operator

$$
\begin{aligned}
&\mathrm{DEL}\colon \mathcal{C}^1(\overline{\Omega}) \to \mathcal{C}^0(\overline{\hat{\Omega}}, \mathbb{R}^d)\\
&\mathrm{DEL}(L_d)(x_0,x_1,x_2) = \nabla_2 L_d(x_0,x_1) + \nabla_1 L_d(x_1,x_2).
\end{aligned} \tag{39}
$$

Here $\nabla_j L_d$ denotes the partial derivatives with respect to the jth input argument of L_d.

Assume that the observations $\hat{\Omega}_0 = \{(x_0^{(j)}, x_1^{(j)}, x_2^{(j)})\}_{j=1}^{\infty}$ correspond to a discrete Lagrangian dynamical system governed by $L_d^{\mathrm{ref}} \in \mathcal{C}^1(\overline{\Omega})$ with globally Lipschitz continuous flow map $g\colon \Omega_a \to \Omega_b$, i.e. $\mathrm{DEL}(L_d^{\mathrm{ref}})(x_0, g(x_0,x_1)) = 0$ for all $(x_0,x_1) \in \Omega_a$ and $g(x_0^{(j)}, x_1^{(j)}) = (x_1^{(j)}, x_2^{(j)})$ for all $j \in \mathbb{N}$.

Lemma 5 *The linear functional* $\Phi^{(\infty)}\colon \mathcal{C}^1(\overline{\Omega}) \to \mathcal{C}^0(\overline{\Omega}_a, \mathbb{R}^d)$ *with*

$$
\Phi^{(\infty)}(L_d)(x_0,x_1) = \mathrm{DEL}(L_d)(x_0,x_1,g(x_0,x_1)) \tag{40}
$$

is bounded. □

Proof Indeed, g extends to a globally Lipschitz continuous map $g\colon \overline{\Omega}_a \to \overline{\Omega}_b$ such that $\Phi^{(\infty)}\colon \mathcal{C}^1(\overline{\Omega}) \to \mathcal{C}^0(\overline{\Omega}_a, \mathbb{R}^d)$ is a well-defined map between Banach spaces defined via (40). Let $\|L_d\|_{\mathcal{C}^1(\overline{\Omega})} \leq 1$. In particular,

$$
\sup_{(x_0,x_1)\in\Omega_a} \|\nabla_2 L_d(x_0,x_1)\| \leq 1 \quad \text{and} \quad \sup_{(x_1,x_2)\in\Omega_b} \|\nabla_2 L_d(x_1,x_2)\| \leq 1. \tag{41}
$$

Therefore, by the triangle inequality

$$
\begin{aligned}
\sup_{(x_0,x_1)\in\Omega_a} \mathrm{DEL}(L_d)(x_0,g(x_0,x_1)) &\leq 1 + \sup_{(x_0,x_1)\in\Omega_a} \|\nabla_2 L_d(g(x_0,x_1))\|\\
&\leq 1 + \sup_{(x_1,x_2)\in\Omega_b} \|\nabla_2 L_d(x_1,x_2)\| \leq 2.
\end{aligned} \tag{42}
$$

■

We can now proceed in direct analogy to the continuous setting (Section 6.1.2) with L replaced by L_d and the functional Φ_N of (32) (normalisation conditions) replaced by the corresponding functional for discrete Lagrangians. The details are provided in the following.

Since for each $\overline{x}$ the evaluation functional $\mathrm{ev}_{\overline{x}}\colon f \mapsto f(\overline{x})$ on $\mathcal{C}^0(\overline{\Omega}_a, \mathbb{R}^d)$ is bounded, the following functions constitute bounded linear functionals for $j \in \mathbb{N}$:

$$
\begin{aligned}
\Phi_j\colon \mathcal{C}^1(\overline{\Omega}) \to \mathbb{R}^d, &\qquad \Phi_j(L_d) = \Phi^{(\infty)}(L_d)(\overline{x}^{(j)})\\
\Phi^{(j)}\colon \mathcal{C}^1(\overline{\Omega}) \to (\mathbb{R}^d)^j, &\qquad \Phi^{(j)} = (\Phi_1, \ldots, \Phi_j).
\end{aligned}
$$

For a reference point $\overline{x}_b \in \Omega$ and for $p_b \in \mathbb{R}^d$, $r_b \in \mathbb{R}$ we define the bounded linear functional

$$
\Phi_N\colon \mathcal{C}^1(\overline{\Omega}) \to \mathbb{R}^{d+1}, \qquad \Phi_N(L) = \left(-\nabla_1 L_d(\overline{x_b}), L_d(\overline{x}_b)\right), \tag{43}
$$

related to our normalisation condition for discrete Lagrangians. We will further use the shorthands $\Phi_b^{(k)} = (\Phi_1, \ldots, \Phi_k, \Phi_N)$ and $\Phi_b^{(\infty)} = (\Phi^{(\infty)}, \Phi_N)$, and define

$$y^{(k)} = (0, \ldots, 0, p_b, r_b) \in (\mathbb{R}^d)^k \times \mathbb{R}^d \times \mathbb{R}$$
$$y^{(\infty)} = (0, p_b, r_b) \in \mathcal{C}^0(\overline{\Omega}, \mathbb{R}^d) \times \mathbb{R}^d \times \mathbb{R}.$$

In analogy to Assumption 3 we consider the following assumption.

Assumption 4 *Assume that there is a Hilbert space U with continuous embedding $U \hookrightarrow \mathcal{C}^1(\overline{\Omega})$ such that*

$$\{L_d \in \mathcal{C}^1(\overline{\Omega}) \,|\, \Phi_b^{(\infty)}(L_d) = y^{(\infty)}\} \cap U \neq \emptyset.$$

In other words, U is assumed to contain a Lagrangian consistent with the normalisation and underlying dynamics.

The affine linear subspaces

$$A^{(j)} = \{L_d \in U \,|\, \Phi_b^{(j)}(L_d) = y^{(j)}\} \quad (j \in \mathbb{N})$$
$$A^{(\infty)} = \{L_d \in U \,|\, \Phi_b^{(\infty)}(L_d) = y^{(\infty)}\}$$

are closed in U and not empty by Assumption 4. Therefore, the following extremisation problems constitute convex optimisation problems on U with unique minima in $A^{(j)}$ or $A^{(\infty)}$, respectively:

$$\begin{aligned} L_{d(j)} &= \arg \min_{L_d \in A^{(j)}} \|L_d\|_U \\ L_{d(\infty)} &= \arg \min_{L_d \in A^{(\infty)}} \|L_d\|_U. \end{aligned} \tag{44}$$

Here $\|\cdot\|_U$ denotes the norm in U.

Proposition 6 *The minima $L_{d(j)}$ converge to $L_{d(\infty)}$ in the norm $\|\cdot\|_U$ and, thus, in $\|\cdot\|_{\mathcal{C}^1(\overline{\Omega})}$.* □

Proof The proof is in complete analogy to Proposition 5. ■

Proof (Theorem 2) An application of Theorem 8 (Appendix A.2) to the components of Φ_b^M considered as elements of the dual to the RKHS U shows that the unique minimisers $L_{d(j)}$ in (44) are the conditional means (24) considered in Theorem 2. Notice that the assumption of Theorem 8 on $y = y_b^M$ is fulfilled, see Proposition 10 (Appendix A.3). Thus, Theorem 2 follows from Proposition 6. ■

6.3. Convergence rates of continuous and discrete Lagrangian models

Let $L_{(M)}$ denote the Lagrangian inferred from M observations as in Theorem 1 and let $L_{(\infty)}$ denote the limit as the observations densely fill a compact set. We analyse how fast the learned equations of motions $\mathrm{EL}(L_{(M)}) = 0$ converge to the true equations of motions $\mathrm{EL}(L_{(\infty)}) = 0$ as the distance between observation data points converges to zero. We will show that the extrapolation error $\|\mathrm{EL}(L_{(M)})(x,\dot{x},\ddot{x})\|$ for $(x,\dot{x},\ddot{x})$ an observation of the true dynamical system can be bounded. The bound tends to zero as h^r, where h relates to the maximal distance between data points and r is related to the smoothness of the true dynamics and the kernel. Provided that the observation data fill the space at least as efficiently as uniform meshes, the bound tends to zero as $M^{-\frac{r}{2d}}$, where M is the number of observation points.

Away from degenerate points, the Euler–Lagrange equations implicitly define an acceleration field that expresses $\ddot{x}$ in terms of $(x,\dot{x})$ such that $\mathrm{EL}(L_{(M)})(x,\dot{x},\ddot{x}) = 0$. Roughly speaking, we will show that away from critical points, the convergence rate of the learned acceleration field to the true acceleration field is h^r (or $M^{-\frac{r}{2d}}$ for uniform meshes) as well. Moreover, analogous statements will be shown for discrete Lagrangian models.

6.3.1. Preliminaries: interpolation and smoothening theory

Our proofs make use of statements from interpolation and smoothening theory [3, 36, 57]. Let us recall notions and results that are relevant in our context.

Definition 1 (fill distance) To $\Omega \subset \mathbb{R}^{d'}$ and a finite subset $\Omega_0 \subset \overline{\Omega}$ we define the *fill distance* h of Ω_0 in Ω as

$$h_{\Omega_0} = \mathrm{dist}(\Omega_0, \overline{\Omega}) = \sup_{\overline{x}_0 \in \Omega_0} \min_{\overline{x} \in \overline{\Omega}} \|\overline{x}_0 - \overline{x}\|.$$

□

The fill distance of Ω_0 in Ω coincides with the Hausdorff distance between the sets Ω_0 and Ω.

Example 3 (Fill distance of uniform mesh and of Halton sequence) When $\overline{\Omega} \subset \mathbb{R}^{d'}$ is a d'-dimensional cube and Ω_0 is a uniform mesh with mesh width $\Delta\overline{x}$ then $h_{\Omega_0} = \sqrt{d'}\Delta\overline{x}/2$. If Ω_0 contains M points,

$$h_{\Omega_0} = \frac{\sqrt{d'}}{2(\sqrt[d']{M} - 1)}.$$

Figure 10 shows the fill distance h_{Ω_0} when Ω_0 is an equidistant uniform mesh on a d'-dimensional cube $\overline{\Omega}$ and when Ω_0 is a Halton sequence with the same number of elements. Here $2h_{\Omega_0}$ corresponds to the maximal distance between any two points in $\Omega_0 \cup \partial\Omega$. It illustrates that in low dimensions Halton sequences reduce the fill distance roughly at a similar rate as uniform meshes. □

In our analysis we will make use of the following theorem from interpolation and smoothening theory.

Theorem 3 (Sobolev bounds) *Let $\Omega \subset \mathbb{R}^{d'}$ be a bounded domain with a Lipschitz continuous boundary. Let $r > \frac{1}{2}d'$. Then there exist constants $\delta_r, C_r > 0$ (depending on Ω and r) such that for any finite $\Omega_0 \subset \overline{\Omega}$ with $h_{\Omega_0} \le \delta_r$, for any $u \in W^r(\Omega)$ with $u|_{\Omega_0} \equiv 0$, and for any $l = 0, \dots, r$*

$$\|u\|_{W^l(\Omega)} \le C_r (h_{\Omega_0})^{r-l} \|u\|_{W^r(\Omega)}.$$

□

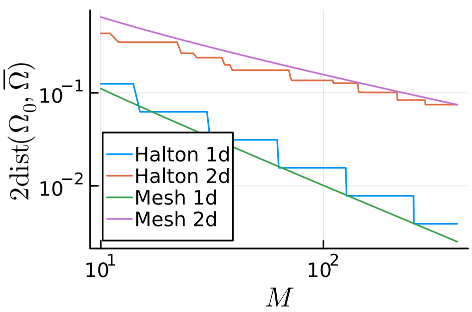

Figure 10: Max. distance between any two points in $\Omega_0 \cup \partial\Omega$ for $\overline{\Omega} = [0,1]^d$, $d = 1, 2$, and Ω_0 a uniform mesh or a Halton sequence with M elements. (See Example 3.)

In the theorem's statement, $W^r(\Omega) = W^{r,2}(\Omega)$ denotes the Sobolev space (defined in Remark 8). u is continuous by the Sobolev embedding theorem [1, §4]. $u|_{\Omega_0}$ denotes the restriction of the function u to the set Ω_0.

Proof This is a special case of [3, Cor. 4.1]. ∎

6.3.2. Convergence rates for continuous Lagrangian models

We will analyse the convergence rates of the inferred equations of motions and the acceleration field to true equations of motions and the true acceleration field as the fill-distance of observations converges to zero. This will, in particular, provide a theoretical explanation of the numerically observed convergence behaviour in Figure 5.

Assumption 5 (Underlying system and RKHS) *Assume $\Omega \subset \mathbb{R}^{2d}$ is open, bounded and has locally Lipschitz boundary. Consider a kernel $K \colon \Omega \times \Omega \to \mathbb{R}$ such that the RKHS U embeds continuously into the Sobolev space[6] $W^{r+2}(\Omega)$ for $r > 2 + d$. Assume that the true acceleration $\ddot{x}$ can be described by a function $g_{\mathrm{ref}} \colon \Omega \to \mathbb{R}^d$ with $g_{\mathrm{ref}} \in (W^r(\Omega))^d$.*

When Assumption 5 holds, then by the Sobolev embedding theorem [1, §4], $W^{r+2}(\Omega)$ embeds continuously into $C^4(\overline{\Omega})$. Therefore, the kernel K necessarily fulfils sufficient smoothness properties such that for $p_b \in \mathbb{R}^d$, $c_b \in \mathbb{R}$, $\overline{x}_b \in \Omega$ we can define to any finite subset $\Omega_0 = \{(x, \dot{x})\}_{j=1}^M \subset \Omega$ a Lagrangian $L_{\Omega_0} \in U$ by (19).

Consider data-driven equations of motions $\mathrm{EL}(L_{\Omega_0})(x, \dot{x}, \ddot{x}) = 0$ inferred from finitely many observations $(x^{(j)}, \dot{x}^{(j)}, \ddot{x}^{(j)})$ with $\Omega_0 = \{(x^{(j)}, \dot{x}^{(j)})\}_j$. The following Theorem provides a bound on the extrapolation error $\mathrm{EL}(L_{\Omega_0})(x, \dot{x}, \ddot{x})$ on observations $(x, \dot{x}, \ddot{x})$ of the true system.

Theorem 4 (Convergence rates for equations of motion) *Let Assumption 5 hold. For $p_b \in \mathbb{R}^d$, $c_b \in \mathbb{R}$, $\overline{x}_b \in \Omega$ assume there exists a Lagrangian $L_{\mathrm{ref}} \in U$ consistent with the normalization (27) and the dynamics, i.e. $\mathrm{EL}(L_{\mathrm{ref}})(\overline{x}, g_{\mathrm{ref}}(\overline{x})) = 0$, for all $\overline{x} \in \overline{\Omega}$.*

[6] See Remark 8 for a definition.

Denote by ${}_k\Phi^\infty(L)(\overline{x}) = {}_k\mathrm{EL}(L)(\overline{x}, g_{\mathrm{ref}}(\overline{x}))$ *for* $L \in U$*,* $\overline{x} \in \Omega$ *the* k*th component of* $\mathrm{EL}(L)(\overline{x}, g_{\mathrm{ref}}(\overline{x}))$ *(*$k = 1, \ldots, d$*).*

Then there exist constants $\delta_r, C_r > 0$ *such that for all finite* $\Omega_0 \subset \overline{\Omega}$ *with* $h_{\Omega_0} = \mathrm{dist}(\Omega_0, \overline{\Omega}) < \delta_r$ *and for all* $l = 0, 1, \ldots, r$*,* $k = 1, \ldots, d$

$$\|{}_k\Phi^\infty(L_{\Omega_0})\|_{W^l(\Omega)} \leq C_r h_{\Omega_0}^{r-l} \|L_{\mathrm{ref}}\|_U.$$

□

Proof All components ${}_k\Phi^\infty$ of the map $\Phi^\infty \colon U \to (W^r(\Omega))^d$ have bounded operator norm: For any $L \in U$ and any $k = 1, \ldots, d$

$${}_k\Phi^{(\infty)}(L) = \sum_{i=1}^{d} \frac{\partial^2 L}{\partial \dot{x}^k \partial \dot{x}^i} \cdot g^i_{\mathrm{ref}} + \frac{\partial^2 L}{\partial x \partial \dot{x}^k} \cdot \dot{x}^i - \frac{\partial L}{\partial x^k}.$$

In the above formula, $\dot{x}^i$ needs to be interpreted as the projection map sending a point $(x, \dot{x}) \in \Omega$ to the component $\dot{x}^i$. Using the triangle inequality and the Cauchy-Schwarz inequality on the Hilbert space $W^r(\Omega)$ we have

$$\begin{aligned}
\|{}_k\Phi^{(\infty)}(L)\|_{W^r(\Omega)} &\leq \sum_{i=1}^{d} \left(\left\| \frac{\partial^2 L}{\partial \dot{x}^k \partial \dot{x}^i} \cdot g^i_{\mathrm{ref}} \right\|_{W^r(\Omega)} + \left\| \frac{\partial^2 L}{\partial x^i \partial \dot{x}^k} \cdot \dot{x}^i \right\|_{W^r(\Omega)} \right) + \left\| \frac{\partial L}{\partial x^k} \right\|_{W^r(\Omega)} \\
&\leq \sum_{i=1}^{d} \left(\left\| \frac{\partial^2 L}{\partial \dot{x}^k \partial \dot{x}^i} \right\|_{W^r(\Omega)} \|g^i_{\mathrm{ref}}\|_{W^r(\Omega)} + \left\| \frac{\partial^2 L}{\partial x^i \partial \dot{x}^k} \right\|_{W^r(\Omega)} \|\dot{x}^i\|_{W^r(\Omega)} \right) \\
&\quad + \left\| \frac{\partial L}{\partial x^k} \right\|_{W^r(\Omega)} \\
&\leq \|L\|_{W^{r+2}(\Omega)} \left(1 + \sum_{i=1}^{d} (\|g^i_{\mathrm{ref}}\|_{W^r(\Omega)} + \|\dot{x}^i\|_{W^r(\Omega)}) \right)
\end{aligned}$$

As the embedding $U \hookrightarrow W^{r+2}(\Omega)$ is continuous, there exists $c_r > 0$ such that $\|L\|_{W^{r+2}(\Omega)} \leq c_r \|L\|_U$. Thus, ${}_k\Phi^\infty \colon U \to W^r(\Omega)$ has bounded operator norm $\|{}_k\Phi^\infty\|_{U, W^r(\Omega)}$.

By Theorem 3 there exist $\delta_r > 0$, $\tilde{C}_r > 0$ such that for all finite $\Omega_0 \subset \overline{\Omega}$ (defining L_{Ω_0}) with $h_{\Omega_0} < \delta_r$ and all $l = 0, \ldots, r$

$$\|{}_k\Phi^\infty(L_{\Omega_0})\|_{W^l(\Omega)} \leq \tilde{C}_r h^{r-l} \|{}_k\Phi^\infty(L_{\Omega_0})\|_{W^r(\Omega)}.$$

As by Remark 4, $L_{\Omega_0} \in U$ minimizes the RKHS-norm while fulfilling the normalisation condition (27) and $\Phi^\infty(L_{\Omega_0})(\overline{x}) = 0$ for all $\overline{x} \in \Omega_0$. As $L_{\mathrm{ref}} \in U$ fulfils (27) and the stricter condition $\Phi^\infty(L_{\mathrm{ref}})(\overline{x}) = 0$ for all $\overline{x} \in \Omega$, we have $\|L_{\Omega_0}\|_U \leq \|L_{\mathrm{ref}}\|_U$. Therefore, combining all estimates, we arrive at

$$\begin{aligned}
\|{}_k\Phi^\infty(L_{\Omega_0})\|_{W^l(\Omega)} &\leq \tilde{C}_r h^{r-l} \|{}_k\Phi^\infty(L_{\Omega_0})\|_{W^r(\Omega)} \\
&\leq \tilde{C}_r h^{r-l} \|{}_k\Phi^\infty\|_{U, W^r(\Omega)} \|L_{\Omega_0}\|_U \\
&\leq \tilde{C}_r h^{r-l} \|{}_k\Phi^\infty\|_{U, W^r(\Omega)} \|L_{\mathrm{ref}}\|_U.
\end{aligned}$$

This proves the claim. ■

As by Example 3, when observations are obtained over a sequence of uniform meshes in $\Omega \subset \mathbb{R}^{2d}$ then the convergence rate predicted in Theorem 4 is $M^{-\frac{r}{2d}}$, where M is the number of observations.

When the dynamics and the kernel are smooth, then Theorem 4 can be applied for any r. However, we expect that the constants δ_r, C_r grow with r. Thus, higher and higher convergence rates become dominant as the fill distance h_{Ω_0} decreases. This is discussed in the following Corollary.

Corollary 2 (Convergence rates equations of motions, Gaussian kernel) *Let $\Omega \subset \mathbb{R}^d$ open, bounded with locally Lipschitz boundary and $K \colon \Omega \times \Omega \to \mathbb{R}$ the squared exponential kernel. Assume the observed acceleration field $g_{\mathrm{ref}} \colon \Omega \to \mathbb{R}^d$ is smooth and all derivatives are bounded on $\overline{\Omega}$. For $p_b \in \mathbb{R}^d$, $c_b \in \mathbb{R}$, $\overline{x}_b \in \Omega$ assume there exists a Lagrangian $L_{\mathrm{ref}} \in U$ consistent with the normalization (27) and the dynamics. Then for all $r \in \mathbb{N}$ there exist $C_r, \delta_r > 0$ such that for all finite subsets $\Omega_0 \subset \Omega$ (defining L_{Ω_0}) with $h_{\Omega_0} < \delta_r$ and for all $l = 0, \dots, r$*

$$\|\overline{x} \mapsto {}_k\mathrm{EL}(L_{\Omega_0})(\overline{x}, g_{\mathrm{ref}}(\overline{x}))\|_{W^l(\Omega)} \leq C_r h_{\Omega_0}^{r-l} \|L_{\mathrm{ref}}\|_U$$

for any component $k = 1, \dots, d$. □

Proof As K is the squared exponential kernel, its reproducing kernel Hilbert space U embeds continuously into any Sobolev space $W^m(\Omega)$ $(m > 1)$ [14, Thm.4.48]. Thus, Assumption 5 is fulfilled for any $r > 2 + d$. Therefore, for $r > 2 + d$ statement follows by Theorem 4. For $r \leq 2 + d$ the statement can be deduced from the statement with $r = 3 + d$ for a sufficiently small $0 < \delta_r < \delta_{3+d}$ and sufficiently large $C_r > C_{3+d}$. ■

For a Lagrangian $L \in \mathcal{C}^2(\Omega)$ at non-degenerate points, i.e. where the matrix $\frac{\partial^2 L}{\partial \dot{x} \partial \dot{x}}$ is invertible, we can define the acceleration field

$$g(L)(\overline{x}) = \left(\frac{\partial^2 L}{\partial \dot{x} \partial \dot{x}}(\overline{x})\right)^{-1} \left(\frac{\partial L}{\partial x}(x, \dot{x}) - \frac{\partial^2 L}{\partial x \partial \dot{x}}(x, \dot{x}) \cdot \dot{x}\right).$$

It fulfils $\mathrm{EL}(L)(\overline{x}, g(L)(\overline{x})) = 0$. We have the following pointwise convergence result of the acceleration field.

Corollary 3 (Convergence rates of acceleration field) *Under the assumptions of Theorem 4, consider a sequence $(\overline{x}^{(j)})_{j=1}^{\infty} \subset \Omega$ defining a dense subset of Ω. Consider the Lagrangians $L_{(j)}$, $L_{(\infty)}$ characterised in Theorem 1. Assume $L_{(\infty)}$ is non-degenerate at $\overline{x} \in \Omega$. Let $\Omega_0^k := \{\overline{x}^{(j)}\}_{j=1}^k$. Then there exist $J \in \mathbb{N}$, $C_r > 0$ such that for all $k > J$*

$$\|g(L_{(k)})(\overline{x}) - g_{\mathrm{ref}}(\overline{x})\| \leq C_r (h_{\Omega_0^k})^r.$$

□

Again, as by Example 3, when the observations are obtained over uniform meshes in $\Omega \subset \mathbb{R}^{2d}$, then the convergence rate predicted in Corollary 3 is $M^{-\frac{r}{2d}}$, where M is the number of samples.

Proof The Lagrangian $L_{(\infty)}$ is non-degenerate at $\overline{x}$, i.e. all eigenvalues of the symmetric matrix $\frac{\partial^2 L_{(\infty)}}{\partial \dot{x}\partial \dot{x}}(\overline{x})$ are non-zero. Let λ be the eigenvalue closest to 0. Since the Lagrangians $L_{(j)}$ converge to $L_{(\infty)}$ in $\|\cdot\|_{\mathcal{C}^2(\overline{\Omega})}$, there exists J_1 such that for all $k > J_1$ the eigenvalue λ_j of $\frac{\partial^2 L_{(j)}}{\partial \dot{x}\partial \dot{x}}(\overline{x})$ closest to zero fulfils $|\lambda_j - \lambda| < \frac{|\lambda|}{2}$. As

$$\Omega_0^1 \subset \Omega_0^2 \subset \ldots \subset \bigcup_{j=1}^{\infty} \Omega_0^j \subset \overline{\Omega}$$

and $\bigcup_{j=1}^{\infty} \Omega_0^j$ is dense in the compact set $\overline{\Omega}$, we have $h_{\Omega_0^j} \to 0$. By Theorem 4 and $U \subset C(\overline{\Omega})$, there exists $J_2 \in \mathbb{N}$, $C > 0$ such that for all $k > J_2$

$$\|\mathrm{EL}(L_{(k)})(\overline{x}, g_{\mathrm{ref}}(\overline{x}))\| \leq C h^r_{\Omega_0^k},$$

where the norm $\|L_{(\infty)}\|_U$ has been absorbed in the constant C. For $k > J := \max(J_1, J_2)$ we have

$$\begin{aligned} C(h_{\Omega_0^k})^r &\geq \|\mathrm{EL}(L_{(k)})(\overline{x}, g_{\mathrm{ref}}(\overline{x})) - \underbrace{\mathrm{EL}(L_{(k)})(\overline{x}, g(L_{(k)})(\overline{x}))}_{=0}\| \\ &= \left\| \frac{\partial^2 L_{(k)}}{\partial \dot{x}\partial \dot{x}}(\overline{x})(g_{\mathrm{ref}}(\overline{x}) - g(L_{(k)})(\overline{x})) \right\| \\ &\geq \frac{|\lambda|}{2} \left\| g_{\mathrm{ref}}(\overline{x}) - g(L_{(k)})(\overline{x}) \right\| \end{aligned}$$

This proves the claim. ∎

The result is consistent with our numerical experiment presented in Figure 5: as the reference Lagrangian and the kernel in the experiment are smooth, Corollary 3 applies for any $r \in \mathbb{N}$. This confirms our observation that as the number of observation points M increases (and h_{Ω_0} shrinks as visualised in Figure 10), higher and higher convergence rates become dominant until round-off errors become dominant.

6.3.3. Convergence rates of discrete Lagrangian models

We now turn to discrete Lagrangian models. For preparation, we prove the following Cauchy-Schwarz-type inequality.

Lemma 6 *Let $\Omega \subset \mathbb{R}^{2d}$ an open, non-empty, bounded domain with Lipschitz boundary. Let $r > d$ and $g \colon \Omega \to \Omega \subset \mathbb{R}^{2d}$ with $g \in (W^r(\Omega))^{2d}$. Then there exists $C_g > 0$ such that for all $f \in W^r(\Omega)$*

$$\|f \circ g\|_{W^{r-d-1}(\Omega)} \leq C_g \|f\|_{W^r(\Omega)}.$$

□

Proof Denote coordinates of Ω by $z^1, \dots, z^{2d}$. Let $f, g \in W^r(\Omega)$ and let $s \leq r-d-1$. For $m > d$ the Sobolev embedding $W^m(\Omega) \subset \mathcal{C}(\overline{\Omega})$ holds. Therefore, the derivatives $\partial^\alpha f = \frac{\partial^{|\alpha|} f}{(z^1)^{\alpha_1} \dots (z^{2d})^{\alpha_{2d}}}$ of f fulfil $\partial^\alpha f \in \mathcal{C}(\overline{\Omega})$ for all multi-indices α with $|\alpha| \leq s$. Moreover, each component of $\partial^\alpha g$ with $|\alpha| \leq s$ lies in $L^2(\overline{\Omega})$.

A multivariate version of the Faá di Bruno formula [26] shows

$$\partial^\alpha (f \circ g) = \sum_\pi (\partial^{\alpha(\pi)} f) \circ g \cdot g_\pi,$$

where π runs through the set of partitions of the unordered $|\alpha|$-tuple (multi-set)

$$\{\underbrace{1, \dots, 1}_{\alpha_1 \text{ times}}, \dots, \underbrace{2d, \dots, 2d}_{\alpha_{2d} \text{ times}}\}$$

and defines multi-indices $\alpha(\pi)$ for derivatives with $|\alpha(\pi)| \leq s$.

The expression g_π consists of products of derivatives of g of order less than or equal to s. For each π the norm $\|g_\pi\|_{L^2(\Omega)}$ can, therefore, be bounded by a repeated application of the Cauchy inequality in $L^2(\Omega)$. Moreover, $\partial^{\alpha(\pi)} f \in \mathcal{C}(\overline{\Omega})$. As $W^{r-i}(\Omega) \subset C(\overline{\Omega})$ for all $i \leq s$, $\|\partial^\alpha (f \circ g)\|_{L^2(\Omega)}$ is bounded in terms of $\|f\|_{W^r(\Omega)}$ and a g dependent constant $C_g > 0$. ■

Proposition 7 *Let $\Omega \subset \mathbb{R}^{2d}$ an open, non-empty, bounded domain with Lipschitz boundary. Let $r > d$ and $g_{\mathrm{ref}} \in (W^r(\Omega))^d$. Consider the map $\Phi^{(\infty)}$ defined by*

$$\Phi^{(\infty)}(L_d)(\overline{x}) = \mathrm{DEL}(L_d)(\overline{x}, g_{\mathrm{ref}}(\overline{x})).$$

The map $\Phi^{(\infty)}$ considered as a linear operator $\Phi^{(\infty)} \colon W^{r+1}(\Omega) \to (W^{r-d-1}(\Omega))^d$ is bounded. □

Proof Let $k \in \{1, \dots, d\}$ and let ${}_k(\cdot)$ denote the kth component of a function. For $(x_0, x_1) \in \Omega$ define $g(x_0, x_1) = (x_1, g_{\mathrm{ref}}(x_0, x_1))$. We have $g \in (W^r(\Omega))^{2d}$. Let $L_d \in W^{r+1}(\Omega)$. Let $f = {}_k(\nabla_1 L_d)$. We have $f \in W^r(\Omega)$.

Now $\|{}_k(\Phi^{(\infty)})(L_d)\|_{W^{r-d-1}}$ can be bounded in terms of $\|L_d\|_{W^{r+1}}$ using Lemma 6:

$$\begin{aligned}
\|{}_k(\Phi^{(\infty)})(L_d)\|_{W^{r-d-1}} &\leq \|{}_k(\nabla_2 L_d)\|_{W^{r-d-1}} + \|f \circ g\|_{W^{r-d-1}} \\
&\leq \|L_d\|_{W^r} + C_g \|f\|_{W^r} \\
&\leq \|L_d\|_{W^{r+1}} + C_g \|L_d\|_{W^{r+1}} = (1 + C_g)\|L_d\|_{W^{r+1}}
\end{aligned}$$

for a g_{ref} dependent constant $C_g > 0$. ■

Assumption 6 *Let $\Omega \subset \mathbb{R}^{2d}$ an open, non-empty, bounded domain with locally Lipschitz boundary. Consider a kernel $K \colon \Omega \times \Omega \to \mathbb{R}$ such that the RKHS U embeds continuously into the Sobolev space $W^{r+1}(\Omega)$ for $r > 2d+1$. Assume that the true discrete dynamical system $(x_0, x_1) \mapsto (x_1, x_2)$ on Ω can be described by a map $g_{\mathrm{ref}} \colon \Omega \to \Omega$ with $g_{\mathrm{ref}} \in (W^r(\Omega))^d$, where $x_2 = g_{\mathrm{ref}}(x_0, x_1)$.*

Remark 11 (Stricter smoothness assumption) Comparing Assumption 6 and Assumption 5, the smoothness assumptions on the dynamics and on the RKHS $\mathcal{U}$ appear to be stricter for discrete Lagrangian models than for continuous models. This is related to the requirement that the target space of $\Phi^{(\infty)}$ (Proposition 7) embeds into $\mathcal{C}(\overline{\Omega})$ to apply smoothening theory (Theorem 3). □

When Assumption 6 holds, then by the Sobolev embedding theorem [1, §4], $W^{r+1}(\Omega)$ embeds continuously into $C^{2+d}(\overline{\Omega}) \subset C^2(\overline{\Omega})$. Therefore, the kernel K necessarily fulfils sufficient smoothness properties such that for $p_b \in \mathbb{R}^d$, $c_b \in \mathbb{R}$, $\overline{x}_b \in \Omega$ we can define to any finite subset $\Omega_0 = \{(x_0, x_1)\}_{j=1}^M \subset \Omega$ a Lagrangian $L_{d,\Omega_0} \in U$ by (24).

The following Theorem provides a bound on the extrapolation error $\mathrm{DEL}(L_{d,\Omega_0})(x_0, x_1, x_2)$ on observations (x_0, x_1, x_2) of the true system, when L_{d,Ω_0} is inferred from finitely many observations. Theorem 5 corresponds to Theorem 4, which relates to continuous Lagrangian models.

Theorem 5 *Under Assumption 6, assume that for $p_b \in \mathbb{R}^d$, $c_b \in \mathbb{R}$, $\overline{x}_b \in \Omega$ there exists a discrete Lagrangian L_d^{ref} consistent with the normalisation* (37) *and the dynamics, i.e.* $\mathrm{DEL}(L_d^{\mathrm{ref}})(\overline{x}, g_{\mathrm{ref}}(\overline{x})) = 0$ *for all $\overline{x} \in \Omega$. Denote by ${}_k\Phi^{\infty}(L_d)(\overline{x}) = {}_k\mathrm{DEL}(L_d)(\overline{x}, g_{\mathrm{ref}}(\overline{x}))$ for $L_d \in U$, $\overline{x} \in \Omega$ the kth component of* $\mathrm{DEL}(L_d)(\overline{x}, g_{\mathrm{ref}}(\overline{x}))$ *($k = 1, \ldots, d$).*

Then there exist constants $\delta_r, C_r > 0$ such that for all finite $\Omega_0 \subset \overline{\Omega}$ (defining L_{d,Ω_0}) with $h_{\Omega_0} = \mathrm{dist}(\Omega_0, \overline{\Omega}) < \delta_r$ and for all $l = 0, 1, \ldots, r-d-1$, $k = 1, \ldots, d$

$$\|{}_k\Phi^{\infty}(L_{d,\Omega_0})\|_{W^l(\Omega)} \leq C_r h_{\Omega_0}^{r-d-1-l} \|L_d^{\mathrm{ref}}\|_U.$$ □

Proof Let $C_{r,\Phi} > 0$ be a bound for the operator norm of $\Phi^{(\infty)} \colon W^{r+1}(\Omega) \to (W^{r-d-1}(\Omega))^d$ (see Proposition 7). As $r > 2d+1$, by Theorem 3 there exists $\delta_r, C_r > 0$ such that for all finite subsets $\Omega_0 \subset \overline{\Omega}$ (defining L_{d,Ω_0}) with $h_{\Omega_0} \leq \delta_r$ and for all $l = 0, \ldots, r-d-1$ we have

$$\begin{aligned}
\|\Phi^{(\infty)}(L_{d,\Omega_0})\|_{W^l(\Omega)} &\leq \tilde{C}_r (h_{\Omega_0})^{r-d-1-l} \|\Phi^{(\infty)}(L_{d,\Omega_0})\|_{W^{r-d-1}(\Omega)} \\
&\leq \tilde{C}_r (h_{\Omega_0})^{r-d-1-l} C_{r,\Phi} \|L_{d,\Omega_0}\|_{W^{r+1}(\Omega)} \\
&\leq \tilde{C}_r (h_{\Omega_0})^{r-d-1-l} C_{r,\Phi} \tilde{c} \|L_{d,\Omega_0}\|_U,
\end{aligned}$$

where $\tilde{c}$ is related to the embedding $U \subset W^{r+1}$. Discrete Lagrangians obtained via (24) fulfil a minimisation principle as explained in the proof of Theorem 2 (in direct analogy to Remark 4, which is formulated for continuous Lagrangians). Thus $\|L_{d,\Omega_0}\|_U \leq \|L_d^{\mathrm{ref}}\|_U$. This completes the proof. ■

For $L_d \in \mathcal{C}^1(\Omega)$, $(x_0^*, x_1^*), (x_1^*, x_2^*) \in \Omega$ with $\mathrm{DEL}(L_d)(x_0^*, x_1^*, x_2^*) = 0$ and $\nabla_{1,2} L_d(x_1^*, x_2^*) = \frac{\partial^2 L_d}{\partial x_1 \partial x_2}(x_1^*, x_2^*)$ invertible, the triple (x_0^*, x_1^*, x_2^*) is called *non-degenerate motion segment* of L_d. By the implicit function theorem we can define a unique continuous map g on a connected open neighbourhood $\mathfrak{O}$ of $(L_d, (x_0^*, x_1^*)) \in \mathcal{C}^1(\Omega) \times \Omega$ with $g(L_d)(x_0^*, x_1^*) = x_2^*$ and

$$\mathrm{DEL}(L_d)(\overline{x}, g(L_d)(\overline{x})) = 0 \quad \forall (L_d, \overline{x}) \in \mathfrak{O}.$$

The map $g(L_d)$ is the *discrete evolution rule* of the discrete dynamical system defined by the Lagrangian L_d.

We have the following pointwise convergence result.

Corollary 4 (Convergence rates discrete evolution rule) *In the setting of Theorem 2 assume $\Omega = \Omega_a = \Omega_b$ and that Assumption 6 is fulfilled in addition. Let $\overline{x}^* = (x_0^*, x_1^*), (x_1^*, x_2^*) \in \Omega$ with $x_2^* = g_{\mathrm{ref}}(\overline{x})$ be a nondegenerate motion sequence of the limit Lagrangian $L_{d,(\infty)}$ defined in Theorem 2. Denote $\Omega_0^k := \{\overline{x}^{(j)}\}_{j=1}^k$. Then there exist $K \in \mathbb{N}$, $C_r > 0$ such that for all $k > K$ the discrete evolution $g(L_{d,(k)})(\overline{x}^*)$ can be defined with $g(L_{d,(k)})(\overline{x}^*) \to x_2^*$ and*

$$\|g(L_{d,(k)})(\overline{x}^*) - g_{\mathrm{ref}}(\overline{x}^*)\| \leq C_r (h_{\Omega_0^k})^{r-d-1}. \qquad \square$$

Proof By Theorem 2, $L_{d,(k)}$ converges to $L_{d,(\infty)}$ in the RKHS U, which is continuously embedded into $\mathcal{C}^2(\overline{\Omega})$ by Assumption 6. Therefore and by the non-degeneracy properties of $L_{d,(\infty)}$, there exists a neighbourhood O of $g_{\mathrm{ref}}(\overline{x}^*)$, an index $K \in \mathbb{N}$, and $\delta > 0$ such that for all $k > K$ and all $\overline{x} \in O$ each row and each column vector of $\nabla_{1,2} L_{d,(k)}(\overline{x}) = \frac{\partial^2 L_{d,(k)}}{\partial x^1 \partial x^2}(\overline{x})$ have norm at least $\delta > 0$. We can assume O to be convex and K so large that for all $k > K$ the line segment between $g(L_{d,(k)})(\overline{x}^*)$ and $g(L_{d,(\infty)})(\overline{x}^*)$ is contained in O.

Let $j \in \{1, \ldots, d\}$ denote an index. Again, we denote the component of a function by a lower-left-aligned index. By Theorem 5 (with $l = 0$) there exists $\tilde{C}_r > 0$ such that for all $k > K$

$$\begin{aligned}
\tilde{C}_r (h_{\Omega_0^k})^{r-d-1} &\geq \|{}_j\mathrm{DEL}(L_{d,(k)})(\overline{x}^*, g_{\mathrm{ref}}(\overline{x}^*))\| \\
&= \|{}_j\mathrm{DEL}(L_{d,(k)})(\overline{x}^*, g_{\mathrm{ref}}(\overline{x}^*)) - \underbrace{{}_j\mathrm{DEL}(L_{d,(k)})(\overline{x}^*, g(L_{d,(k)})(\overline{x}^*))}_{=0}\| \\
&= \|{}_j\nabla_1 L_{d,(k)}(x_1^*, g_{\mathrm{ref}}(\overline{x}^*)) - {}_j\nabla_1 L_{d,(k)}(x_1^*, g(L_{d,(k)})(\overline{x}^*))\| \\
&= \|\nabla_2({}_j\nabla_1) L_{d,(k)}(x_1^*, x')^\top (g_{\mathrm{ref}}(\overline{x}^*) - g(L_{d,(k)})(\overline{x}^*))\| \\
&\geq \delta \|g_{\mathrm{ref}}(\overline{x}^*) - g(L_{d,(k)})(\overline{x}^*))\|.
\end{aligned}$$

Above, x' lies on the line segment between $g(L_{d,(k)})(\overline{x}^*)$ and $g(L_{d,(\infty)})(\overline{x}^*)$. Its existence is guaranteed by the intermediate value theorem. The expression $\nabla_2({}_j\nabla_1)L_{d,(k)}$ denotes the gradient of ${}_j\nabla_1 L_{d,(k)}$ with respect to the second input slot of $L_{d,(k)}$. The last inequality holds true since the norm of each row and each column of $\nabla_{1,2} L_{d,(k)}(\overline{x}^*, x')$ is bounded from below by $\delta > 0$. Thus the theorem follows with $C_r = \frac{\tilde{C}_r}{\delta}$. ■

7. Summary

We have introduced a method to learn general continuous Lagrangians and discrete Lagrangians from observational data of dynamical systems that are governed by variational ordinary differential equations. The method is based on kernel-based, meshless collocation methods for solving partial differential equations [52]. In our context, collocation methods are used to solve the Euler–Lagrange equations that we interpret as a partial differential equations for a Lagrangian function L, or discrete Lagrangian L_d, respectively. Additionally, the use of Gaussian processes gives access to a statistical framework that allows for a quantification of the model uncertainty of the identified dynamical system. This could be used for adaptive sampling of data points. Uncertainty quantification can be efficiently computed for any quantity that is linear in the Lagrangian, such as the Hamiltonian or symplectic structure of the system, which is of relevance in the context of system identification. We prove the convergence of the methods to a true Lagrangian and prove convergence rates for the inferred equations of motion, acceleration fields, and evolution rules as the maximal distance of observation data points converges to zero.

The article overcomes the major difficulty that Lagrangians are not uniquely determined by a system's motions and the presence of degenerate solutions to the Euler–Lagrange equations. This is tackled by a careful consideration of regularisation conditions that reduce the gauge freedom of Lagrangians but do not restrict the generality of the ansatz. Our method profits from implicit regularisation that can be understood as an extremisation of a reproducing kernel Hilbert space norm, based on techniques of game theory [44]. This interpretation as convex optimisation problems is the key ingredient that allows us to provide a rigorous proof of convergence of the method as the maximal distance of observation data points converges to zero.

In [38] we have extended the method to dynamical systems governed by variational partial differential equations. Another direction of research is to adapt the method to dynamical systems with low regularity such as systems with collisions and to incorporate noise models into our statistical framework. Furthermore, a combination with detection methods for Lie group variational symmetries [18, 30] or with detection methods of travelling waves [40, 42] is of interest. This may allow for a quantitative analysis of the interplay of symmetry assumptions and model uncertainty.

Acknowledgments

The author acknowledges the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen and computing time provided by the Paderborn Center for Parallel Computing (PC2).

Data availability

The data that support the findings of this study are openly available in the GitHub repository Christian-Offen/Lagrangian_GP at
`https://github.com/Christian-Offen/Lagrangian_GP`. An archived version [39] of release v1.0 of the GitHub repository is openly available at `https://doi.org/10.5281/zenodo.11093645`.

Appendices

A. Gaussian fields

A.1. Definitions

We recall from [44] definitions and properties of Gaussian fields and their interpretation as weak random variables.

Definition 2 Let V be a topological vector space and V^* its topological dual. A linear operator $T\colon V^* \to V$ is *positive symmetric* if $\psi(T\phi) = \phi(T\psi)$ for all $\phi, \psi \in V^*$ and $\phi(T\phi) \geq 0$ for all $\phi \in V^*$. □

Let $(B, \|\cdot\|_B)$ be a separable Banach space with quadratic norm $\|\cdot\|_B$, i.e. there exists a linear, positive symmetric, bijection $Q\colon B^* \to B$ such that $\|u\|_B = (Q^{-1}u)(u)$. Even though this implies that B is a Hilbert space, the Banach space terminology is used as the dual pairing of B^* and B does not coincide with the inner product pairing via the Riesz representation theorem. Moreover, as any positive symmetric linear operator $B^* \to B$ is automatically continuous [44, Prop. 11.2], $Q\colon B^* \to B$ is continuous.

Definition 3 ([44, Def. 17.3]) Let $T\colon B^* \to B$ be a positive symmetric linear operator, $u \in B$, $(\mathcal{A}, \Sigma, \mathbb{P})$ a probability space with $\mathbb{P}$ a Borel measure, and $H \subset L^2(\mathcal{A}, \Sigma, \mathbb{P})$ a linear subspace such that each $X \in H$ is a Gaussian random variable. A linear map

$$\xi\colon B^* \to H \subset L^2(\mathcal{A}, \Sigma, \mathbb{P})$$

is a *Gaussian field with mean u and covariance operator T* if for each $\phi \in B^*$ the random variable $\xi(\phi)$ is normally distributed with mean $\phi(u)$ and covariance $\phi(T\phi)$, i.e. $\xi(\phi) \sim \mathcal{N}(\phi(u), \phi(T\phi))$. We denote such a field by $\xi \sim \mathcal{N}(u, T)$. When $u = 0$, then we say ξ is a *centred Gaussian field.* □

Remark 12 (Notation) Consider a Gaussian field $\xi \sim \mathcal{N}(u, T)$, $\xi\colon B^* \to L^2(\mathcal{A}, \Sigma, \mathbb{P})$ as in Definition 3. The Gaussian field ξ post-composed with evaluation at $\omega \in \mathcal{A}$ is a linear map $\xi(\cdot)(\omega)\colon B^* \to \mathbb{R}$, which is an element in the algebraic dual to B^*. Strictly speaking, the map $\omega \mapsto \xi(\cdot)(\omega)$ cannot be interpreted as a B-valued random variable because it takes values in the *algebraic dual* to B^* but not necessarily in the topological dual $B^{**} \cong B$ because $\xi\colon B^* \to L^2(\mathcal{A}, \Sigma, \mathbb{P})$ might not be bounded. However, $\omega \mapsto \xi(\cdot)(\omega)$ admits the interpretation as a *weak* B-valued random variable [44, §17.4] and we say that ξ is a *Gaussian field on B*.

For $\phi \in B^*$ we define $\phi(\xi) := \xi(\phi)$, which is the notation used in Sections 4 to 6. □

Theorem 6 ([44, Thm. 17.4]) *To any $u \in B$ and symmetric positive covariance operator T a Gaussian field $\xi \sim \mathcal{N}(u, T)$ exists.* □

Lemma 7 *Let $\xi \sim \mathcal{N}(u, T)$ for $u \in B$ and a positive symmetric operator $T\colon B^* \to B$. Then for $\phi, \psi \in B^*$ the covariance of $\xi(\phi)$ and $\xi(\psi)$ is given as*

$$\operatorname{cov}(\xi(\psi), \xi(\phi)) = \mathbb{E}[(\xi(\psi) - \psi(u))(\xi(\phi) - \phi(u))] = \psi T \phi.$$ □

Proof As covariances are invariant under shifts, without loss of generality we may assume $u = 0$. We have

$$\begin{aligned}(\psi + \phi)T(\psi + \phi) &= \mathrm{cov}(\xi(\psi + \phi), \xi(\psi + \phi)) = \mathbb{E}[\xi(\psi + \phi)\xi(\psi + \phi)] \\ &= \mathbb{E}[\xi(\psi)\xi(\psi) + 2\xi(\psi)\xi(\phi) + \xi(\phi)\xi(\phi)] \\ &= \psi T\psi + 2\mathrm{cov}(\xi(\psi), \xi(\phi)) + \phi T\phi\end{aligned}$$

It follows that $\mathrm{cov}(\xi(\psi), \xi(\phi)) = \psi T\phi$. ∎

A.2. Conditional expectation and variance

Let $\xi \sim \mathcal{N}(u, T)$ be a Gaussian field with covariance operator T and let $\phi, \phi_1, \ldots, \phi_m \in B^*$. Let $\Phi = (\phi_1, \ldots, \phi_m)$ and denote $\xi(\Phi) = (\xi(\phi_1), \ldots, \xi(\phi_m))$, $\Phi(u) = (\phi_1(u), \ldots, \phi_m(u))$, $\Theta = (\phi_i T\phi_j)_{i,j=1}^m \in \mathbb{R}^{m\times m}$, $\Theta_0 = (\phi T\phi_j)_{j=1}^m \in \mathbb{R}^m$, $\Theta_0^0 = \phi T\phi$. Using Lemma 7, the joint distribution of $(\xi(\phi), \xi(\Phi))\colon \mathcal{A} \to \mathbb{R}^{m+1}$ is given as

$$\begin{pmatrix}\xi(\phi)\\ \xi(\Phi)\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\phi(u)\\ \Phi(u)\end{pmatrix}, \begin{pmatrix}\Theta_0^0 & \Theta_0^\top\\ \Theta_0 & \Theta\end{pmatrix}\right).$$

We have $\xi(\Phi) - \Phi(u) \in \mathrm{range}(\Theta)$ almost surely [20, Prop. 2.7]. Here $\mathrm{range}(\Theta)$ denotes the span of the columns of Θ. Let $y \in \Phi(u) + \mathrm{range}(\Theta)$. Let the expression $\Theta^\dagger$ denote the Penrose pseudo-inverse of Θ. Using $\Theta_0^\top = \Theta_0^\top\Theta^\dagger\Theta$ [20, Prop.2.16], the two linear systems of equations

$$\Theta z = y - \Phi(u) \quad \text{and} \quad \Theta Z = \Theta_0, \tag{45}$$

are solvable.

Proposition 8 ([20, Prop. 3.13]) *The conditional distribution of $\xi(\phi)$ given $\xi(\Phi) = y$ is given as*

$$\xi(\phi)|\xi(\Phi) = y \sim \mathcal{N}(\phi(u) + \Theta_0^\top\Theta^\dagger(y - \Phi(u)), \Theta_0^0 - \Theta_0^\top\Theta^\dagger\Theta_0).$$ □

Remark 13 The expressions $\Theta^\dagger(y - \Phi(u))$ and $\Theta^\dagger\Theta_0$ denote the (column-wise) least square solutions to (45). However, since $\mathrm{null}(\Theta) \subseteq \mathrm{null}(\Theta_0^\top)$ [20, Prop. 2.16] any solution to (45) will yield the same conditional distribution. □

Remark 14 As by the existence result (Theorem 6), the function $\phi \mapsto \xi(\phi)|\xi(\Phi) = y$ can be interpreted as a Gaussian field with $\overline{\xi} \sim \mathcal{N}(\overline{u}, \overline{T})$ with mean

$$\overline{u} = u + (T\Phi)^\top\Theta^\dagger(y - \Phi(u)) \in B, \text{ with } T\Phi = (T\phi_1, \ldots, T\phi_m)$$

and covariance given by the positive symmetric operator $\overline{T}\colon B^* \to B$ (in the sense of Definition 2)

$$\overline{T} = T - (T\Phi)^\top\Theta^\dagger(T\Phi).$$

For an interpretation of $\overline{\xi}$ as an orthogonal projection of ξ and a measure theoretic discussion, we refer to [44]. □

The following statements are helpful to characterize conditional means of Gaussian fields by an extremization principle.

Theorem 7 ([44, Thm. 12.5]) *Let $\phi_1, \ldots, \phi_m \in B^*$ be linearly independent. Define $\Theta \in \mathbb{R}^{m\times m}$ by it elements $\Theta_{i,j} = \phi_i(Q\phi_j)$. Denote $\Phi = \big(\phi_1, \ldots, \phi_m\big)^\top \in (B^*)^m$ and $Q\Phi = \big(Q\phi_1, \ldots, Q\phi_m\big)^\top \in B^m$. Then Θ is invertible and for any $y \in \mathbb{R}^m$*

$$\Psi = y^\top \Theta^{-1} Q\Phi = \sum_{i,j=1}^{m} y_i (\Theta^{-1})_{i,j} Q\phi_j$$

is the minimizer of the convex optimization problem

$$\Psi = \arg \min_{\{\Psi \in B \,|\, \Phi(\Psi) = y\}} \|\Psi\|_B.$$

□

We weaken the assumptions of Theorem 7 slightly.

Theorem 8 *Let $\phi_1, \ldots, \phi_m \in B^*$. Define $\Theta \in \mathbb{R}^{m\times m}$ by $\Theta_{i,j} = \phi_i(Q\phi_j)$. Denote $\Phi = \big(\phi_1, \ldots, \phi_m\big)^\top \in (B^*)^m$ and $Q\Phi = \big(Q\phi_1, \ldots, Q\phi_m\big)^\top \in B^m$. For any $y \in \operatorname{range}(\Phi\colon B \to \mathbb{R}^m)$*

$$\Psi = y^\top \Theta^{\dagger} Q\Phi = \sum_{i,j=1}^{m} y_i (\Theta^{\dagger})_{i,j} Q\phi_j$$

is the minimizer of the convex optimization problem

$$\Psi = \arg \min_{\{\Psi \in B \,|\, \Phi(\Psi) = y\}} \|\Psi\|_B.$$

□

Proof In preparation of the argument we first proof the following Lemma

Lemma 8 *We have*

$$\ker(\Theta\colon \mathbb{R}^m \to \mathbb{R}^m) \subseteq \ker(\mathbb{R}^m \to B, x \mapsto x^\top Q\Phi).$$

□

Proof The proof is inspired by [20, Prop. 2.16]. Let $\phi \in B^*$. As the bijection Q is positive symmetric, the following matrix is symmetric, positive semi-definite

$$\Sigma = \begin{pmatrix} \phi Q \phi & \phi(Q\Phi)^\top \\ \Phi^\top Q\phi & \Phi^\top Q\Phi \end{pmatrix} \in \mathbb{R}^{(m+1)\times(m+1)}$$

Therefore, for any $x \in \ker \Theta = \ker \Phi^\top Q\Phi$, $\alpha, \beta \in \mathbb{R}$ we have

$$0 \le \begin{pmatrix} \beta & \alpha x^\top \end{pmatrix} \Sigma \begin{pmatrix} \beta \\ \alpha x \end{pmatrix} = \beta^2 \phi Q\phi + 2\alpha\beta\phi(Q\Phi)^\top x$$

As this holds for all $\alpha, \beta \in \mathbb{R}$ we conclude $\phi((Q\Phi)^\top x) = 0$. Since B is a Hilbert space and $\phi((Q\Phi)^\top x) = 0$ holds for all $\phi \in B^*$ we conclude $(Q\Phi)^\top x = 0$. ■

Let $\{\tilde{\phi}_j\}_{j=1}^{\tilde{m}} \subset B^*$ be a basis for the linear span $\operatorname{span}\{\phi_j\}_{j=1}^{m}$ ($\tilde{m} \leq m$). The basis elements define the vector $\tilde{\Phi} = \begin{pmatrix}\tilde{\phi}_1, & \dots, & \tilde{\phi}_{\tilde{m}}\end{pmatrix}^\top \in (B^*)^{\tilde{m}}$. By linear independence of $\tilde{\phi}_j$, for $k = 1, \dots, m$ there exist unique $\alpha_{kj} \in \mathbb{R}$ ($j = 1, \dots, \tilde{m}$) with

$$\phi_k = \sum_{j=1}^{\tilde{m}} \alpha_{kj} \tilde{\phi}_j.$$

This defines a unique matrix $A = (\alpha_{ij}) \in \mathbb{R}^{m \times \tilde{m}}$ with $\Phi = A\tilde{\Phi}$ with linearly independent columns. Moreover, the matrix $\tilde{\Theta} \in \mathbb{R}^{\tilde{m} \times \tilde{m}}$ defined by $\tilde{\Theta}_{i,j} = \tilde{\phi}_i(Q(\tilde{\phi}_j))$ is invertible.

Since $\tilde{\Phi} \colon B \to \mathbb{R}^{\tilde{m}}$ is surjective,

$$\operatorname{range}(\Phi \colon B \to \mathbb{R}^m) = \operatorname{range}(A \circ \tilde{\Phi} \colon B \to \mathbb{R}^m) = \operatorname{range}(A \colon \mathbb{R}^{\tilde{m}} \to \mathbb{R}^m).$$

Let $y \in \operatorname{range}(\Phi) = \operatorname{range}(A)$. Define $\tilde{y} = A^\dagger y$ and $\tilde{z} = \tilde{\Theta}^{-1}\tilde{y}$. As A has linearly independent columns, it is an isomorphism onto $\operatorname{range}(A)$ such that for any $\Psi \in B$ we have

$$\begin{aligned} (A \circ \tilde{\Phi})(\Psi) = \Phi(\Psi) = y = A\tilde{y} \\ \iff \tilde{\Phi}(\Psi) = \tilde{y}. \end{aligned}$$

Thus, the following minima coincide

$$\arg \min_{\{\Psi \in B \,|\, \Phi(\Psi) = y\}} \|\Psi\|_B = \arg \min_{\{\Psi \in B \,|\, \tilde{\Phi}(\Psi) = \tilde{y}\}} \|\Psi\|_B.$$

By Theorem 7 this minimum coincides with $\tilde{\Psi} = \tilde{z}^\top Q \tilde{\Phi}$. To complete the proof of the theorem, it remains to prove the following Lemma.

Lemma 9 *Consider the linear system $\Theta z = y$ for $z \in \mathbb{R}^m$. Then $\Theta z = y$ is solvable and for any solution z*

$$\Psi = z^\top Q \Phi \quad \textit{and} \quad \tilde{\Psi} = \tilde{z}^\top Q \tilde{\Phi}$$

coincide. □

Proof Let $\tilde{z}$ be the solution to $\tilde{\Theta}\tilde{z} = \tilde{y}$ and set $\tilde{\Psi} = \tilde{z}^\top Q \tilde{\Phi}$. Using linearity of Q, we have

$$\tilde{\Psi} = \tilde{z}^\top Q \tilde{\Phi} = \tilde{z}^\top Q A^\dagger \Phi = ((A^\dagger)^\top \tilde{z})^\top Q \Phi = \overline{z}^\top Q \Phi$$

with $\overline{z} := (A^\dagger)^\top \tilde{z}$. We have

$$A^\dagger \Theta \overline{z} = A^\dagger \Theta (A^\dagger)^\top \tilde{z} = \tilde{\Theta}\tilde{z} = \tilde{y} = A^\dagger y.$$

As $A^\dagger A = \mathrm{Id}_{\tilde{m}}$, the restriction $A^\dagger|_{\operatorname{range}(A)} \colon \operatorname{range}(A) \to \mathbb{R}^{\tilde{m}}$ is an isomorphism. Therefore, as $y \in \operatorname{range}(A)$ it follows that $\overline{z} = (A^\dagger)^\top \tilde{z}$ solves the linear system

$$\Theta z = y. \tag{46}$$

For any other solution $z \in \mathbb{R}^m$ of (46) it holds that

$$z - \overline{z} \in \ker \Theta \subseteq \ker((\Theta\Phi)^\top : \mathbb{R}^m \to \mathbb{R}), \qquad \blacksquare$$

where the inclusion holds by Lemma 8. Therefore, $\tilde{\Psi} = \tilde{z}^\top Q\Phi = \overline{z}^\top Q\Phi = z^\top Q\Phi = \Psi$. $\blacksquare$

This completes the proof of Theorem 8.

A.3. Applicability of Proposition 8 in Section 4

To apply Proposition 8 in Section 4.1.3, we need to verify $y_b^M \in \text{range}(\Theta)$.

Proposition 9 *Employing notation of Section 4, Assumption 1 implies* $y_b^M \in \text{range}(\Theta)$. $\square$

Proof Denote the components of Φ_b^M by $\phi_1, \dots, \phi_{\overline{M}} \in U^*$, where $\overline{M} = Md + d + 1$. Let $\{\tilde{\phi}_j\}_{j=1}^{\tilde{M}} \subset U^*$ be a basis for the linear span $\text{span}\{\phi_j\}_{j=1}^{\overline{M}}$ ($\tilde{M} \le \overline{M}$). The basis elements define the vector $\tilde{\Phi} = \begin{pmatrix} \tilde{\phi}_1, & \dots, & \tilde{\phi}_{\tilde{M}} \end{pmatrix}^\top \in (U^*)^{\tilde{M}}$. By the linear independence of $\tilde{\phi}_j$, for $k = 1, \dots, \overline{M}$ there exist unique $\alpha_{kj} \in \mathbb{R}$ ($j = 1, \dots, \tilde{M}$) with

$$\phi_k = \sum_{j=1}^{\tilde{M}} \alpha_{kj} \tilde{\phi}_j.$$

This defines a unique matrix $A = (\alpha_{ij}) \in \mathbb{R}^{\overline{M} \times \tilde{M}}$ with $\Phi_b^M = A\tilde{\Phi}$ with linearly independent columns. Define $\tilde{\Theta} \in \mathbb{R}^{\tilde{M} \times \tilde{M}}$ by $\tilde{\Theta}_{i,j} = \tilde{\phi}_i(\mathcal{K}(\tilde{\phi}_j))$. Recall that $\Theta_{i,j} = \phi_i(\mathcal{K}(\phi_j))$ defines $\Theta \in \mathbb{R}^{\overline{M} \times \overline{M}}$.

Lemma 10 *We have*

$$\Theta = A\tilde{\Theta}A^\top. \qquad \square$$

Proof Using linearity of $\mathcal{K} : U^* \to U$,

$$\begin{aligned} \Theta_{i,j} = \phi_i(\mathcal{K}(\phi_j)) &= \sum_{k=1}^{\tilde{M}} \alpha_{ik} \tilde{\phi}_k \left(\mathcal{K} \left(\sum_{s=1}^{\tilde{M}} \alpha_{js} \tilde{\phi}_s \right) \right) \\ &= \sum_{k,s=1}^{\tilde{M}} \alpha_{ik} \alpha_{js} \tilde{\phi}_k (\mathcal{K}(\tilde{\phi}_s)) = \sum_{k,s=1}^{\tilde{M}} \alpha_{ik} \tilde{\Theta}_{k,s} \alpha_{js}. \end{aligned} \qquad \blacksquare$$

The matrix $\tilde{\Theta}$ is invertible by construction. (Indeed, we could have chosen $\tilde{\phi}_j$ such that $\tilde{\Theta}$ is the identity matrix.) Moreover, A is injective such that $A^\top : \mathbb{R}^{\overline{M}} \to \mathbb{R}^{\tilde{M}}$ is surjective. Therefore, by Lemma 10, $\text{range}(\Theta) = \text{range}(A)$. Viewing $\Phi_b^M : U \to \mathbb{R}^{\overline{M}}$, $\tilde{\Phi} : U \to \mathbb{R}^{\tilde{M}}$, $A : \mathbb{R}^{\tilde{M}} \to \mathbb{R}^{\overline{M}}$ as linear maps,

$$\text{range}(\Phi_b^M : U \to \mathbb{R}^{\overline{M}}) = \text{range}(A \circ \tilde{\Phi} : U \to \mathbb{R}^{\overline{M}}) \subset \text{range}(A) = \text{range}(\Theta). \qquad \blacksquare$$

Assumption 1 implies $y_b^M \in \text{range}(\Phi_b^M)$. Therefore, $y_b^M \in \text{range}(\Theta)$.

In the setting of discrete Lagrangians, an application of Proposition 8 is justified as by the following Proposition.

Proposition 10 *Employing notation of Section 4.2, Assumption 2 implies* $y_b^M \in \mathrm{range}(\Theta)$. □

Proof The proof follows in complete analogy to Proposition 9. ■

B. Alternative regularisation

The following proposition justifies an alternative regularisation strategy. As it involves non-linear conditions, we prefer the regularisation strategy presented in the main body of the document. However, it is presented here for comparison with regularisation strategies for learning of Lagrangian densities using neural networks [42].

Proposition 11 *Let* $\overline{x}_b = (x_b, \dot{x}_b) \in T\mathbb{R}^d \cong \mathbb{R}^d \times \mathbb{R}^d$ *and* $\mathring{L}$ *a Lagrangian with* $\frac{\partial \mathring{L}}{\partial \dot{x} \partial \dot{x}}(\overline{x}_b)$ *non-degenerate. Let* $c_b \in \mathbb{R}$, $p_b \in \mathbb{R}^d$, $c_\omega > 0$. *There exists a Lagrangian* L *such that* L *is equivalent to* $\mathring{L}$ *and*

$$L(\overline{x}_b) = c_b, \quad \mathrm{Mm}(L)(\overline{x}_b) = \frac{\partial L}{\partial \dot{x}}(\overline{x}_b) = p_b, \quad N_\omega(L)(\overline{x}_b) = \left| \det \left(\frac{\partial^2 L}{\partial \dot{x} \partial \dot{x}}(\overline{x}_b) \right) \right| = c_\omega. \tag{47}$$

□

Proof Let $\mathring{c}_b = \mathring{L}(\overline{x}_b)$, $\mathring{p}_b = \mathrm{Mm}(\mathring{L})(\overline{x}_b)$, $\mathring{c}_\omega = N_\omega(\mathring{L})(\overline{x}_b)$. The quantity $\mathring{c}_\omega$ is not zero since $\frac{\partial \mathring{L}}{\partial \dot{x} \partial \dot{x}}(\overline{x}_b)$ is non-degenerate. We set

$$\rho = \sqrt[d]{\left| \frac{c_\omega}{\mathring{c}_\omega} \right|}, \quad F(x) = x^\top (p_b - \rho \mathring{p}_b), \quad c = c_b - \dot{x}_b^\top (p_b - \rho \mathring{p}_b) - \rho \mathring{c}_b.$$

Now the Lagrangian $L = \rho \mathring{L} + \mathrm{d}_t F + c$ is equivalent to $\mathring{L}$ and fulfils (14). ■

The condition $N_\omega(L)(\overline{x}_b) = c_\omega > 0$ may be compared to the regularisation strategies for training Lagrangians modelled as neural networks in [42]: denoting observation data by $\hat{x}^{(j)} = (x^{(j)}, \dot{x}^{(j)}, \ddot{x}^{(j)})$, in [42] (transferred to our continuous ode setting) parametrises L as a neural network and considers the minimisation of a loss function function $\ell = \ell_{\mathrm{data}} + \ell_{\mathrm{reg}}$ with data consistency term

$$\ell_{\mathrm{data}} = \sum_j \| \mathrm{EL}(L)(\hat{x}^{(j)}) \|^2$$

and with regularisation term ℓ_{reg} that maximises the regularity of the Lagrangian at data points $\hat{x}^{(j)} = (x^{(j)}, \dot{x}^{(j)}, \ddot{x}^{(j)})$

$$\ell_{\mathrm{reg}} = \sum \left\| \left(\frac{\partial^2 L}{\partial \dot{x} \partial \dot{x}}(x^{(j)}, \dot{x}^{(j)}) \right)^{-1} \right\|.$$

The corresponding statement for discrete Lagrangians is as follows.

Proposition 12 *Let* $\overline{x}_b = (x_{0b}, x_{1b}) \in \mathbb{R}^d \times \mathbb{R}^d$ *and* $\mathring{L}_d$ *a discrete Lagrangian with* $\mathrm{Mm}^-(\overline{x}_b)$ *non-degenerate. Let* $c_b \in \mathbb{R}$, $p_b \in \mathbb{R}^d$, $c_\omega > 0$. *There exists a discrete Lagrangian* L_d *such that* L_d *is equivalent to* $\mathring{L}_d$ *and*

$$L_d(\overline{x}_b) = c_b, \quad \mathrm{Mm}^-(L_d)(\overline{x}_b) = p_b, \quad N_\omega^-(L_d)(\overline{x}_b) = \left|\det\left(\frac{\partial^2 L_d}{\partial x_0 \partial x_1}(\overline{x}_b)\right)\right| = c_\omega. \tag{48}$$

□

Proof Let $\mathring{c}_b = \mathring{L}_d(\overline{x}_b)$, $\mathring{p}_b = \mathrm{Mm}^-(\mathring{L}_d)(\overline{x}_b)$, $\mathring{c}_\omega = N_\omega^-(\mathring{L}_d)(\overline{x}_b)$. The quantity $\mathring{c}_\omega$ is not zero since $\frac{\partial \mathring{L}_d}{\partial x_0 \partial x_1}(\overline{x}_b)$ is non-degenerate. We set

$$\rho = \sqrt[d]{\left|\frac{c_\omega}{\mathring{c}_\omega}\right|}, \quad F(x) = x^\top(p_b - \rho\mathring{p}_b), \quad c = c_b - \rho\mathring{c}_b - (x_{1b} - x_{0b})^\top(p_b - \rho\mathring{p}_b).$$

Now the Lagrangian $L_d = \rho\mathring{L}_d + \Delta_t F + c$ is equivalent to L_d and fulfils (48). ■

Again, the condition $N_\omega^-(L)(\overline{x}_b) = c_\omega > 0$ may be compared to the regularisation strategies for training discrete Lagrangians modelled as neural networks in [42]: denoting observation data by $\hat{x}^{(j)} = (x_0^{(j)}, x_1^{(j)}, x_2^{(j)})$, in [42] (when transferred to our discrete ode setting) parametrises L_d as a neural network and considers the minimisation of a loss function function $\ell = \ell_{\mathrm{data}} + \ell_{\mathrm{reg}}$ with data consistency term

$$\ell_{\mathrm{data}} = \sum_j \|\mathrm{DEL}(L_d)(\hat{x}^{(j)})\|^2$$

and with regularisation term ℓ_{reg} that maximises the regularity of the Lagrangian at data points $\hat{x}^{(j)} = (x_0^{(j)}, x_1^{(j)}, x_2^{(j)})$:

$$\ell_{\mathrm{reg}} = \sum \left\| \left(\frac{\partial^2 L}{\partial x_0 \partial x_1}(x_0^{(j)}, x_1^{(j)})\right)^{-1} \right\|.$$

C. Derivation of symplectic structure induced by discrete Lagrangians

Denote the coordinate of the domain of definition $\mathbb{R}^d \times \mathbb{R}^d$ of a discrete Lagrangian L_d by (x_0, x_1). Consider the two discrete Legendre transforms $\Phi^\pm \colon \mathbb{R}^d \times \mathbb{R}^d \to T^*\mathbb{R}^d$ [35] with

$$\Phi^-(x_0, x_1) = (x_0, -\frac{\partial L}{\partial x_0}(x_0, x_1)) \qquad \Phi^+(x_0, x_1) = (x_1, \frac{\partial L}{\partial x_1}(x_0, x_1)).$$

When we pullback the canonical symplectic structure $\sum_{k=1}^d \mathrm{d}q^k \wedge \mathrm{d}p_k$ on $T^*\mathbb{R}^d$ to the discrete phase space $\mathbb{R}^d \times \mathbb{R}^d$ with $\Phi^\pm$ we obtain

$$\begin{aligned}
\mathrm{Sympl}^-(L_d) &= \sum_{s=1}^{d} \mathrm{d}x_0^s \wedge \mathrm{d}\left(-\frac{\partial L_d}{\partial x_0^s}\right) = \sum_{r,s=1}^{d} -\frac{\partial^2 L_d}{\partial x_0^s \partial x_0^r} \mathrm{d}x_0^s \wedge \mathrm{d}x_0^r - \frac{\partial^2 L_d}{\partial x_0^s \partial x_1^r} \mathrm{d}x_0^s \wedge \mathrm{d}x_1^r \\
&= \sum_{r,s=1}^{d} -\frac{\partial^2 L_d}{\partial x_0^s \partial x_1^r} \mathrm{d}x_0^s \wedge \mathrm{d}x_1^r \\
\mathrm{Sympl}^+(L_d) &= \sum_{s=1}^{d} \mathrm{d}x_1^s \wedge \mathrm{d}\left(\frac{\partial L_d}{\partial x_1^s}\right) = \sum_{r,s=1}^{d} \frac{\partial^2 L_d}{\partial x_1^s \partial x_0^r} \mathrm{d}x_1^s \wedge \mathrm{d}x_0^r + \frac{\partial^2 L_d}{\partial x_1^s \partial x_1^r} \mathrm{d}x_1^s \wedge \mathrm{d}x_1^r \\
&= \sum_{r,s=1}^{d} \frac{\partial^2 L_d}{\partial x_1^s \partial x_0^r} \mathrm{d}x_1^s \wedge \mathrm{d}x_0^r
\end{aligned}$$

We see $\mathrm{Sympl}^-(L_d) = \mathrm{Sympl}^+(L_d)$.

The 2-form corresponds to the notion of a *discrete Lagrangian symplectic form* in [35, §1.3.2].

References

[1] Robert A. Adams and John J.F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Elsevier, 2003. doi:10.1016/S0079-8169(03)80006-5.

[2] Takehiro Aoshima, Takashi Matsubara, and Takaharu Yaguchi. Deep discrete-time lagrangian mechanics. *ICLR SimDL*, 5 2021. URL: https://simdl.github.io/files/49.pdf.

[3] Rémi Arcangéli, María Cruz López de Silanes, and Juan José Torrens. An extension of a bound for functions in Sobolev spaces, with applications to (m, s)-spline interpolation and smoothing. *Numerische Mathematik*, 107(2):181–211, Aug 2007. doi:10.1007/s00211-007-0092-z.

[4] Vladimir I. Arnold. *Introduction to perturbation theory*, pages 271–300. Springer New York, New York, NY, 1989. URL: http://dx.doi.org/10.1007/978-1-4757-2063-1_10, doi:10.1007/978-1-4757-2063-1_10.

[5] Victor Barwell and Alan George. A comparison of algorithms for solving symmetric indefinite systems of linear equations. *ACM Trans. Math. Softw.*, 2(3):242–251, September 1976. doi:10.1145/355694.355697.

[6] Pau Batlle, Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. Error analysis of kernel/gp methods for nonlinear and parametric pdes, 2023. arXiv:2305.04962.

[7] Tom Bertalan, Felix Dietrich, Igor Mezić , and Ioannis G. Kevrekidis. On learning Hamiltonian systems from data. *Chaos*, 29(12):121107, dec 2019. doi:10.1063/1.5128231.

[8] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. doi:10.1137/141000671.

[9] James R. Bunch and Linda Kaufman. Some stable methods for calculating inertia and solving symmetric linear systems. *Mathematics of Computation*, 31(137):163–179, 1977. doi:10.2307/2005787.

[10] José F Carinena and Luis A Ibort. Non-Noether constants of motion. *Journal of Physics A: Mathematical and General*, 16(1):1, 1 1983. doi:10.1088/0305-4470/16/1/010.

[11] Renyi Chen and Molei Tao. Data-driven prediction of general Hamiltonian dynamics via learning exactly-symplectic maps. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1717–1727. PMLR, 18–24 Jul 2021. URL: https://proceedings.mlr.press/v139/chen21r.html, arXiv:2103.05632.

[12] Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M. Stuart. Solving and learning nonlinear pdes with Gaussian processes. *Journal of Computational Physics*, 447:110668, 2021. doi:10.1016/j.jcp.2021.110668.

[13] Yuhan Chen, Baige Xu, Takashi Matsubara, and Takaharu Yaguchi. Variational principle and variational integrators for neural symplectic forms. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023. URL: https://openreview.net/forum?id=XvbJqbW3rf.

[14] Andreas Christmann and Ingo Steinwart. *Kernels and Reproducing Kernel Hilbert Spaces*, pages 110–163. Springer New York, New York, NY, 2008. doi:10.1007/978-0-387-77242-4_4.

[15] Nathaël Da Costa, Marvin Pförtner, Lancelot Da Costa, and Philipp Hennig. Sample path regularity of Gaussian processes from the covariance kernel, 2024. arXiv:2312.14886.

[16] Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks, 2020. doi:10.48550/ARXIV.2003.04630.

[17] Marco David and Florian Méhats. Symplectic learning for Hamiltonian neural networks. *Journal of Computational Physics*, 494:112495, 2023. doi:10.1016/j.jcp.2023.112495.

[18] Eva Dierkes, Christian Offen, Sina Ober-Blöbaum, and Kathrin Flaßkamp. Hamiltonian neural networks with automatic symmetry detection. *Chaos*, 33(6):063115, 06 2023. 063115. doi:10.1063/5.0142969.

[19] Joseph Diestel. *Sequences and Series in Banach Spaces*. Springer New York, 1984. doi:10.1007/978-1-4612-5200-9.

[20] Morris L. Eaton. *Multivariate Statistics*. John Wiley & Sons, 1 2007. doi:10.1214/lnms/1196285102.

[21] Sølve Eidnes and Kjetil Olsen Lye. Pseudo-Hamiltonian neural networks for learning partial differential equations. *Journal of Computational Physics*, 500:112738, 2024. doi:10.1016/j.jcp.2023.112738.

[22] Giulio Evangelisti and Sandra Hirche. Physically consistent learning of conservative lagrangian systems with gaussian processes. In *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022. doi:10.1109/CDC51059.2022.9993123.

[23] Jinchao Feng, Charles Kulick, Yunxiang Ren, and Sui Tang. Learning particle swarming models from data with Gaussian processes. *Mathematics of Computation*, 93(349):2391–2437, 11 2023. doi:10.1090/mcom/3915.

[24] I.M. Gelfand, S.V. Fomin, and R.A. Silverman. *Calculus of Variations*. Dover Books on Mathematics. Dover Publications, 2000.

[25] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian Neural Networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/26cd8ecadce0d4efd6cc8a8725cbd1f8-Paper.pdf, arXiv:1906.01563.

[26] Michael Hardy. Combinatorics of partial derivatives. *The Electronic Journal of Combinatorics*, 13(1):R1, 2006. doi:10.37236/1027.

[27] Marc Henneaux. Equations of motion, commutation relations and ambiguities in the Lagrangian formalism. *Annals of Physics*, 140(1):45–64, 1982. doi:10.1016/0003-4916(82)90334-7.

[28] Jianyu Hu, Juan-Pablo Ortega, and Daiying Yin. A structure-preserving kernel method for learning Hamiltonian systems, 2024. arXiv:2403.10070.

[29] Pengzhan Jin, Zhen Zhang, Aiqing Zhu, Yifa Tang, and George Em Karniadakis. SympNets: Intrinsic structure-preserving symplectic networks for identifying Hamiltonian systems. *Neural Networks*, 132:166–179, 2020. doi:10.1016/j.neunet.2020.08.017.

[30] Yana Lishkova, Paul Scherer, Steffen Ridderbusch, Mateja Jamnik, Pietro Liò, Sina Ober-Blöbaum, and Christian Offen. Discrete Lagrangian neural networks with automatic symmetry discovery. *IFAC-PapersOnLine*, 56(2):3203–3210, 2023. 22nd IFAC World Congress. doi:10.1016/j.ifacol.2023.10.1457.

[31] Fei Lu, Ming Zhong, Sui Tang, and Mauro Maggioni. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proceedings of the National Academy of Sciences*, 116(29):14424–14433, 6 2019. doi:10.1073/pnas.1822012116.

[32] Giuseppe Marmo and Giuseppe Morandi. On the inverse problem with symmetries, and the appearance of cohomologies in classical Lagrangian dynamics. *Reports on Mathematical Physics*, 28(3):389–410, 1989. doi:10.1016/0034-4877(89)90071-2.

[33] Giuseppe Marmo and Claudio Rubano. On the uniqueness of the Lagrangian description for charged particles in external magnetic field. *Il Nuovo Cimento A*, 98(4):387–399, 10 1987. doi:10.1007/bf02902083.

[34] Jerrold E. Marsden and Tudor S. Ratiu. *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems*. Springer New York, New York, NY, 1999. doi:10.1007/978-0-387-21792-5.

[35] Jerrold E. Marsden and Matthew West. Discrete mechanics and variational integrators. *Acta Numerica*, 10:357–514, 2001. doi:10.1017/S096249290100006X.

[36] Francis J. Narcowich, Joseph D. Ward, and Holger Wendland. Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Mathematics of Computation*, 74(250):743–763, 2005. URL: http://www.jstor.org/stable/4100087.

[37] Sina Ober-Blöbaum and Christian Offen. Variational learning of Euler–Lagrange dynamics from data. *Journal of Computational and Applied Mathematics*, 421:114780, 2023. doi:10.1016/j.cam.2022.114780.

[38] Christian Offen. Machine learning of discrete field theories with guaranteed convergence and uncertainty quantification, 2024. arXiv:2407.07642.

[39] Christian Offen. Software: Christian-Offen/Lagrangian_GP: Initial release of GitHub Repository, 4 2024. doi:10.5281/zenodo.11093645.

[40] Christian Offen and Sina Ober-Blöbaum. Learning discrete Lagrangians for variational pdes from data and detection of travelling waves. In Frank Nielsen and Frédéric Barbaresco, editors, *Geometric Science of Information*, volume 14071, pages 569–579, Cham, 2023. Springer Nature Switzerland. doi:10.1007/978-3-031-38271-0_57.

[41] Christian Offen and Sina Ober-Blöbaum. Symplectic integration of learned Hamiltonian systems. *Chaos*, 32(1):013122, 1 2022. doi:10.1063/5.0065913.

[42] Christian Offen and Sina Ober-Blöbaum. Learning of discrete models of variational PDEs from data. *Chaos*, 34:013104, 1 2024. doi:10.1063/5.0172287.

[43] Juan-Pablo Ortega and Daiying Yin. Learnability of linear port-Hamiltonian systems, 2023. arXiv:2303.15779.

[44] Houman Owhadi and Clint Scovel. *Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2019. doi:10.1017/9781108594967.

[45] Marvin Pförtner, Ingo Steinwart, Philipp Hennig, and Jonathan Wenger. Physics-informed Gaussian process regression generalizes linear pde solvers, 2024. arXiv:2212.12474.

[46] Hong Qin. Machine learning and serving of discrete field theories. *Scientific Reports*, 10(1), 11 2020. doi:10.1038/s41598-020-76301-0.

[47] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(65):1939–1959, 2005. URL: http://jmlr.org/papers/v6/quinonero-candela05a.html.

[48] Christopher Rackauckas and Qing Nie. Differentialequations.jl–a performant and feature-rich ecosystem for solving differential equations in julia. *Journal of Open Research Software*, 5(1):15, 2017. doi:10.5334/jors.151.

[49] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005. doi:10.7551/mitpress/3206.001.0001.

[50] Katharina Rath, Christopher G. Albert, Bernd Bischl, and Udo von Toussaint. Symplectic Gaussian process regression of maps in Hamiltonian systems. *Chaos*, 31(5):053121, 05 2021. doi:10.1063/5.0048129.

[51] Tomáš Roubíček. *Calculus of Variations*, pages 1–38. John Wiley & Sons, Ltd, 2015. doi:10.1002/3527600434.eap735.

[52] Robert Schaback and Holger Wendland. Kernel techniques: From machine learning to meshless methods. *Acta Numerica*, 15:543–639, 2006. doi:10.1017/S0962492906270016.

[53] Florian Schäfer, Matthias Katzfuss, and Houman Owhadi. Sparse Cholesky factorization by Kullback–Leibler minimization. *SIAM Journal on Scientific Computing*, 43(3):A2019–A2046, 2021. doi:10.1137/20M1336254.

[54] Florian Schäfer, T. J. Sullivan, and Houman Owhadi. Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. *Multiscale Modeling & Simulation*, 19(2):688–730, 2021. doi:10.1137/19M129526X.

[55] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In David van Dyk and Max Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL: https://proceedings.mlr.press/v5/titsias09a.html.

[56] Mats Vermeeren. Modified equations for variational integrators. *Numerische Mathematik*, 137(4):1001–1037, 6 2017. doi:10.1007/s00211-017-0896-4.

[57] Holger Wendland and Christian Rieger. Approximate interpolation with applications to selecting smoothing parameters. *Numerische Mathematik*, 101(4):729–748, Oct 2005. doi:10.1007/s00211-005-0637-y.