


# 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

April 30, 2024

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.

**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 or port-Hamiltonian systems from data by approximating the Hamiltonian, pseudo-, or port-Hamiltonian structure by neural networks or Gaussian processes [16, 13, 3, 29, 27, 10, 18]. Additionally, Lie group symmetries are identified in [11]. Alternatively, the symplectic flow map of Hamiltonian systems can be approximated [34, 5, 19] and symplectic structure is identified in [3, 7].

**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 [9, 24, 14, 20]. 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

$$S(x) = \int_{t_0}^{t_1} L(x(t), \dot{x}(t)) dt, \quad (1)$$

where  $x: [t_0, t_1] \rightarrow \mathbb{R}^d$  is a curve with derivative denoted by  $\dot{x}$ . The function  $L$  is a *Lagrangian*. A function  $x: [t_0, t_1] \rightarrow \mathbb{R}^d$  is a solution or *motion* if the action  $S$  is stationary at  $x$  for all variations  $\delta x: [t_0, t_1] \rightarrow \mathbb{R}^d$  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

$$\text{EL}(L)(x(t), \dot{x}(t), \ddot{x}(t)) = 0, \quad t \in (t_0, t_1) \quad (2)$$

with

$$\text{EL}(L) = \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right) - \frac{\partial L}{\partial x} = \frac{\partial^2 L}{\partial x \partial \dot{x}} \ddot{x} + \frac{\partial^2 L}{\partial x \partial x} \dot{x} - \frac{\partial L}{\partial x}. \quad (3)$$

Here,  $\frac{\partial^2 L}{\partial x \partial \dot{x}} = \left( \frac{\partial^2 L}{\partial x^k \partial \dot{x}^l} \right)_{k,l=1}^d$ ,  $\frac{\partial^2 L}{\partial x \partial x} = \left( \frac{\partial^2 L}{\partial x^k \partial x^l} \right)_{k,l=1}^d$  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 [15, 35], for instance.

In the data-driven context,  $L$  is sought as a function of  $\bar{x} = (x, \dot{x})$  such that (3) is fulfilled at observed data points  $\mathcal{D} = \{(x, \dot{x}, \ddot{x})\}_{j=1}^M$ . Once  $L$  is known,  $\text{EL}(x) = 0$  can be solved with a numerical method such as a variational integrator [23].

**Discrete Lagrangian data-driven models** Instead of learning continuous variational principles, in [31] Qin proposes to learn discrete Lagrangian theories by approximating discrete Lagrangians. In discrete Lagrangian theories, motions  $x(t)$  are described at discrete, equidistant times  $t^0 < t^1 < \dots < t^N$  by a sequence of snapshots  $\mathbf{x} = (x_k)_{k=0}^N \subset$

$\mathbb{R}^d$ . The motions constitute stationary points of a discrete action functional

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

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

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

with

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

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 [23].

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 [23, §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 [24] the author provides a method based on Vermeeren’s variational backward error analysis [38] 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 [24]. In [28, 26] 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 processes.

**Novelty** The article

1. introduces a method to learn continuous and discrete Lagrangians from data based on Gaussian process regression with a rigorous proof of its convergence as the distance between data points converges to zero.
2. Moreover, the article systematically discusses the ambiguity of Lagrangians and normalisation strategies for kernel-based learning methods for Lagrangians.

3. Furthermore, the article provides a statistical framework that allows for uncertainty quantification of any linear observable of the dynamical system, such as Hamiltonian functions (energy) or symplectic structure, for instance.

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 [14], an explicit mechanical ansatz is used [2]. In other works regularisation is done implicitly without discussion [9], ad hoc as in the author’s prior work [24], or relates to neural networks [20, 28, 26] only.

Methodologically, the method of the present article stands in the context of meshless collocation methods [36] 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 [30] and was employed to solve well-posed partial differential equations using Gaussian Processes in [6].

**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 review reproducing kernel Hilbert spaces and Gaussian processes. Then we introduce 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 contains theorems that guarantee convergence of the method for Lagrangians and discrete Lagrangians. 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 [15, 35] for details.

We consider the Hamiltonian to a Lagrangian defined via

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

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

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

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

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

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

$$\text{Vol}(L) = \frac{1}{d!} (\text{Sympl}(L))^d = \det \left( \frac{\partial^2 L}{\partial \dot{x}^r \partial \dot{x}^s} \right) dx^1 \wedge d\dot{x}^1 \wedge \dots \wedge dx^d \wedge d\dot{x}^d. \quad (9)$$

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

### 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 [17, 22, 21]. 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.

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

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

$$\tilde{L} - \rho L - c = d_t F$$

---

<sup>1</sup> $\text{Sympl}(L)$  is the pull-back of the canonical symplectic form  $\sum_{s=1}^d dq^s \wedge dp_s$  under the Legendre transform  $T\mathbb{R}^d \rightarrow T^*\mathbb{R}^d$ ,  $(x, \dot{x}) \mapsto (q, p) = (x, \frac{\partial L}{\partial \dot{x}}(x, \dot{x}))$ .

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

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

(See, e.g. [15].) 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.

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

for all regular curves  $x: [t_0, t_1] \rightarrow \mathbb{R}^d$  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 1 (Affine linear motions)** For any twice differentiable  $g: \mathbb{R}^d \rightarrow \mathbb{R}$  with nowhere degenerate Hessian matrix  $\text{Hess}(g)$ , the Lagrangian  $L(x, \dot{x}) = g(\dot{x})$  describes affine linear motions in  $\mathbb{R}^d$ :

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

In general, the existence of alternative Lagrangian densities is related to additional geometric structure and conserved quantities of the system [17, 22, 21, 4]. 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  $F: \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $\rho \in \mathbb{R}$ ,  $c \in \mathbb{R}$ , and  $\tilde{L} = \rho L + d_t F + c$ . We have*

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

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

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

**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  $L_d: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  following definitions in discrete variational calculus [23]. The quantities

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

relate to discrete conjugate momenta at time  $t_j$ . On motions  $\mathbf{x} = (x_k)_{k=0}^N$  that fulfil (4),  $\text{Mm}^-(x_k, x_{k+1})$  and  $\text{Mm}^+(x_{k-1}, x_k)$  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

$$\text{Sympl}(L_d) = \sum_{r,s=1}^d \frac{\partial^2 L_d}{\partial x_1^s \partial x_0^r} dx_1^s \wedge dx_0^r \quad (12)$$

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

$$\text{Vol}(L_d) = \det \left( \frac{\partial^2 L_d}{\partial x_1 \partial x_0} \right) dx_1^1 \wedge dx_0^1 \wedge \dots \wedge dx_1^d \wedge dx_0^d. \quad (13)$$

When  $\frac{\partial^2 L_d}{\partial x_1 \partial x_0}$  is non-degenerate everywhere, then  $\text{Sympl}(L_d)$  is a symplectic form and  $\text{Vol}(L_d)$  its induced volume form on the discrete phase space  $\mathbb{R}^d \times \mathbb{R}^d$ .  $\text{Sympl}(L_d)$  is called *discrete Lagrangian symplectic form* in [23, §1.3.2]. (For consistency with the continuous theory Section 2.1.1 our sign convention differs from [23, §1.3.2]. A derivation can be found in Appendix B.)

### 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

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

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 Lagrangians*.

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

**Lemma 2** *Let  $L_d$  be a Lagrangian depending on  $(x_0, x_1)$ . Consider a continuously differentiable  $F: \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $\rho \in \mathbb{R}$ ,  $c \in \mathbb{R}$ , 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

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

Here  $\nabla F$  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,  $\text{Mm}^\pm$ , SympL, Vol are obtained by a direct computation. The assertion about invariance of non-degenerate points follows from the transformation rule of Vol. ■

### 3. Normalisation of Lagrangians

In the machine learning framework that we will introduce in Section 4, we will employ normalisation 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  $\text{EL}(L) \equiv 0$ . 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 normalisation conditions are covered by the ambiguities presented in Section 2.
- We provide a geometric interpretation of the normalisation conditions and how much geometric structure they fix.

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

**Proposition 1** *Let  $\bar{x}_b = (x_b, \dot{x}_b) \in T\mathbb{R}^d \cong \mathbb{R}^d \times \mathbb{R}^d$ ,  $\hat{x}_\tau = (x_\tau, \dot{x}_\tau, \ddot{x}_\tau) \in (\mathbb{R}^d)^3$  and  $\mathring{L}$  a Lagrangian with  $\text{EL}(\mathring{L})(\hat{x}_\tau) \neq 0$ . 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*

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

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



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

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

Now the Lagrangian  $L = \rho \mathring{L} + d_t F + c$  is equivalent to  $\mathring{L}$  and fulfils (14).  $\blacksquare$

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 1 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*

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

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

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

$$L = \rho \mathring{L} + d_t F + c.$$

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

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

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

**Remark 1** Under genericity assumptions on the dynamics with  $d \geq 2$ , no alternative Lagrangians exist [17]. If a generic dynamical system is governed by a non-degenerate Lagrangian, then any Lagrangian  $L$  with  $\text{EL}(L) = 0$  on all motions that is non-degenerate anywhere, is non-degenerate everywhere.  $\square$

Refer to Proposition 7 of Appendix A for an alternative normalisation strategy for Lagrangians based on normalising symplectic volume. It is comparable to techniques developed in [28] for neural network models of Lagrangians.

The following Proposition implies that 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  $\bar{x}_b = (x_b, \dot{x}_b) \in T\mathbb{R}^d$ ,  $1 \leq k \leq d$ ,  $c_b \in \mathbb{R}$ ,  $p_b \in \mathbb{R}^d$ ,  $c_\tau \in \mathbb{R} \setminus \{0\}$ . Then for any Lagrangian  $L$  equivalent to  $\mathring{L}$  we have*

$$\text{Ham}(L) = \text{Ham}(\mathring{L}), \quad \text{Sym}(L) = \text{Sym}(\mathring{L}).$$

$\square$

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

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

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

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

**Proposition 4** *Let  $\bar{x}_b = (x_{0b}, x_{1b}) \in (\mathbb{R}^d)^2$ ,  $\hat{x}_\tau = (x_{0\tau}, x_{1\tau}, x_{2\tau}) \in (\mathbb{R}^d)^3$  and  $\mathring{L}_d$  a discrete Lagrangian with  $\text{DEL}(L_d)(\hat{x}_b) \neq 0$ . 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*

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

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

PROOF Let  $\mathring{c}_b = \mathring{L}_d(\bar{x}_b)$ ,  $\mathring{p}_b = \text{Mm}^+(\mathring{L}_d)(\bar{x}_b)$ ,  $\mathring{c}_\tau = (\text{DEL}(\mathring{L}_d)(\hat{x}_b))_k$ . 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  $\text{Mm}^-$  replacing  $\text{Mm}^+$ . 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 8 in Appendix A, where it is also compared to regularisation strategies in the neural network context of [28]. □

## 4. Data-driven method

### 4.1. Gaussian processes for 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  $L: \Omega \rightarrow \mathbb{R}$  based on observations  $\hat{x} = (\bar{x}, \ddot{x}) = (x, \dot{x}, \ddot{x}) \in \Omega \times \mathbb{R}^d$  for which  $\text{EL}(L)(\hat{x}) = 0$  on all observations  $\hat{x}$  such that the dynamics (2) to  $L$  approximate the dynamics of an unknown true Lagrangian  $L_{\text{ref}}: \Omega \rightarrow \mathbb{R}$ . The Lagrangian  $L$  will be obtained as the conditional mean of a Gaussian process with guaranteed convergence in the infinite data limit against a true Lagrangian of the motion.

#### 4.1.1. RKHS set-up and Gaussian process

We consider the following set-up that makes use of the theory of reproducing kernel Hilbert spaces (RKHS). Refer to [8, 30] for background material.

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

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

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

with inner product defined as the linear extension of

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

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

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

The map  $\mathcal{K}: U^* \rightarrow U$  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 process*  $\xi \in \mathcal{N}(0, \mathcal{K})$  on  $U$ , i.e.  $\xi$  is a random variable  $\xi: \mathcal{A} \rightarrow U$  on a probability space  $\mathcal{A}$

- with zero mean  $\mathbb{E}(\xi) = 0 \in U$
- such that for any finite collection  $\Phi = (\Phi_1, \dots, \Phi_n)$  with  $\Phi_j \in U^*$  for  $1 \leq j \leq n$ , the random variable  $\Phi(\xi): \mathcal{A} \rightarrow \mathbb{R}^n$  is multivariate-normally distributed  $\Phi(\xi) \in \mathcal{N}(0, \kappa)$  with covariance matrix given as  $\kappa = (\Phi_i(\mathcal{K}(\Phi_j)))_{i,j=1}^n$ .

#### 4.1.2. Data

Assume we observe distinct data points  $\hat{x}^{(j)} = (\bar{x}^{(j)}, \ddot{x}^{(j)}) = (x^{(j)}, \dot{x}^{(j)}, \ddot{x}^{(j)}) \in \Omega \times \mathbb{R}^d$ ,  $j = 1, \dots, M$  of Lagrangian motions. Define  $\text{EL}_{\hat{x}^{(j)}}: U \rightarrow \mathbb{R}^d$  as

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

for  $1 \leq j \leq M$ . Furthermore, let  $\bar{x}_b = (x_b, \dot{x}_b) \in \Omega$  and consider  $\text{Mm}_{\bar{x}_b}: U \rightarrow \mathbb{R}^d$  as

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

Moreover, let  $\text{ev}_{\bar{x}_b}: U \rightarrow \mathbb{R}$  with

$$\text{ev}_{\bar{x}_b}(L) = L(\bar{x}_b)$$

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

$$\Phi_b^M = (\text{EL}_{\hat{x}^{(1)}}, \dots, \text{EL}_{\hat{x}^{(M)}}, \text{Mm}_{\bar{x}_b}, \text{ev}_{\bar{x}_b}). \quad (17)$$

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

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

*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  $\text{Mm}(L)(\bar{x}_b) = p_b$ ,  $L(\bar{x}_b) = c_b$ . The condition  $(\text{EL}(L)(\bar{x}_b))_k = c_\tau$  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 Process

By general theory [30, Cor. 17.12], the posterior distribution of the canonical Gaussian process  $\xi$  conditioned on the linear constraint  $\Phi_b^M(L) = y_b^M$  is again a Gaussian process  $\xi_M = \mathcal{N}(L, \mathcal{K}_{\Phi_b^M})$ . It is fully 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 positive definite 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  $k$ th or  $l$ th component of  $\Phi_b^M$ , respectively. In block matrix form,  $\Theta$  can be written as

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

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

$$\text{EL}_{\hat{x}^{(j)}}^1 \text{EL}_{\hat{x}^{(i)}}^2 K = \text{EL}_{\hat{x}^{(j)}}(\bar{x} \mapsto \text{EL}_{\hat{x}^{(i)}}(\bar{y} \mapsto K(\bar{x}, \bar{y}))) \in \mathbb{R}$$

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

$$\mathcal{K}\Phi_b^M(\bar{x}) = (\text{EL}_{\hat{x}^{(1)}}K(\cdot, \bar{x}), \dots, \text{EL}_{\hat{x}^{(M)}}K(\cdot, \bar{x}), \text{Mm}_{\bar{x}_b}K(\cdot, \bar{x}), K(\bar{x}_b, \bar{x}))^\top$$

the conditional mean  $L$  of the posterior process  $\xi_M$  is given as

$$L = y_b^M{}^\top \theta^{-1} \mathcal{K}\Phi_b^M. \quad (19)$$

The conditional covariance operator  $\mathcal{K}_{\Phi_b^M} : U^* \rightarrow U$  is given by

$$\psi \mathcal{K}_{\Phi_b^M} \phi = \psi \mathcal{K} \phi - (\psi \mathcal{K} \Phi_b^{M\top}) \theta^{-1} (\Phi_b^M \mathcal{K} \phi) \quad (20)$$

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

$$\begin{aligned} \psi \mathcal{K}_{\Phi_b^M} \phi &= \psi^1 \phi^2 K \\ \psi \mathcal{K} \Phi_b^{M\top} &= (\psi^1 \text{EL}_{\hat{x}^{(2)}}^2 K, \dots, \psi^1 \text{EL}_{\hat{x}^{(n)}}^2 K, \psi^1 \text{Mm}_{\bar{x}_b}^2 K, \psi^1 K(\cdot, \bar{x}_b)) \\ \Phi_b^M \mathcal{K} \phi &= (\text{EL}_{\hat{x}^{(2)}}^1 \phi^2 K, \dots, \text{EL}_{\hat{x}^{(n)}}^1 \phi^2 K, \text{Mm}_{\bar{x}_b}^1 \phi^2 K, \phi^2 K(\bar{x}_b, \cdot))^\top. \end{aligned}$$

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.

**Remark 3 (Computational efficiency)** For efficient approximations of solutions of linear systems involving  $\theta$  we refer to the literature on computational aspects of Gaussian process regression, see, for instance, [32, 37].  $\square$

**Remark 4 (Equivalent minimisation problem)** As by general theory [30, Thm 12.5] (also see [6, Prop.2.2]), under appropriate assumptions on the reproducing kernel Hilbert space  $U$  to  $K$  (see Section 6), the conditional mean  $L$  of (19) can alternatively be characterised as the minimiser of the following convex optimisation problem

$$L = \arg \min_{\tilde{L} \in U, \Phi_b^M(\tilde{L}) = y_b^M} \|\tilde{L}\|_U, \quad (21)$$

where  $\|\tilde{L}\|_U$  denotes the reproducing kernel Hilbert space norm. This will play an important role in the convergence proof in Section 6. Besides the exploit for convergence proofs, formulation (21) 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 A using techniques of [6].  $\square$

**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  $(\text{EL}_{(\hat{x}_\tau)})_k$  to  $\Phi_b^M$  for  $\hat{x}_\tau = (x_\tau, \dot{x}_\tau, \ddot{x}_\tau)$  that is not a motion and  $k \in \{1, \dots, d\}$ . While it is realistic to assume knowledge of a data point  $\hat{x}_\tau$  that is not a motion (e.g.  $\hat{x} = (\bar{x}^{(1)}, \ddot{x}^{(1)} + 1)$  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 A discusses an alternative normalisation based on symplectic volume forms. It can be compared to approaches to learn Lagrangians with neural networks [28].  $\square$

#### 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  $\text{Ham}(\xi|_{\Phi_b^M(L)=y_b^M})(\bar{x})$  for  $\bar{x} \in \Omega$  and for  $\text{EL}(\xi|_{\Phi_b^M(L)=y_b^M})(\hat{x}(t))$ , where  $\hat{x} = (x, \dot{x}, \ddot{x})$  is a motion of the approximate system to  $L$ .

## 4.2. Gaussian Processes 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.

In the setting of discrete Lagrangians,  $\Omega \subset \mathbb{R}^d \times \mathbb{R}^d$  is an open, bounded subset containing elements denoted by  $\bar{x} = (x_0, x_1)$ . Observed data corresponds to a collection of 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  $L_d: \Omega \rightarrow \mathbb{R}$  such that discrete motions that fulfil the discrete Euler-Lagrange equations  $\text{DEL}(L_d) = 0$  approximate true motions. With the reinterpretation of  $\Omega$  and of training data points  $\hat{x}^{(j)}$  we can follow the framework for continuous Lagrangians replacing EL by DEL and Mm by  $\text{Mm}^-$  (or  $\text{Mm}^+$ ). In particular, this leads to

$$\Phi_b^M = (\text{DEL}_{\hat{x}^{(1)}}, \dots, \text{DEL}_{\hat{x}^{(M)}}, \text{Mm}^-_{\bar{x}_b}, \text{ev}_{\bar{x}_b}).$$

(cf. (17)) and

$$\Theta = \begin{pmatrix} (\text{DEL}_{\hat{x}^{(j)}}^1 \text{DEL}_{\hat{x}^{(i)}}^2 K)_{ij} & (\text{DEL}_{\hat{x}^{(j)}}^1 \text{Mm}^-_{\bar{x}_b} K)_j & (\text{DEL}_{\hat{x}^{(j)}}^1 \text{ev}_{\bar{x}_b}^2 K)_j \\ (\text{Mm}^-_{\bar{x}_b} \text{DEL}_{\hat{x}^{(i)}}^2 K)_i & \text{Mm}^-_{\bar{x}_b} \text{Mm}^-_{\bar{x}_b} K & \text{Mm}^-_{\bar{x}_b} \text{ev}_{\bar{x}_b}^2 K \\ (\text{ev}_{\bar{x}_b}^1 \text{DEL}_{\hat{x}^{(i)}}^2 K)_i & \text{ev}_{\bar{x}_b}^1 \text{Mm}^-_{\bar{x}_b} K & K(\bar{x}_b, \bar{x}_b). \end{pmatrix} \quad (22)$$

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

$$L_d = y_b^M{}^\top \theta^{-1} \mathcal{K}_{\Phi_b^M} \quad (23)$$

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

$$\psi \mathcal{K}_{\Phi_b^M} \phi = \psi \mathcal{K} \phi - (\psi \mathcal{K}_{\Phi_b^M}{}^\top) \theta^{-1} (\Phi_b^M \mathcal{K} \phi). \quad (24)$$

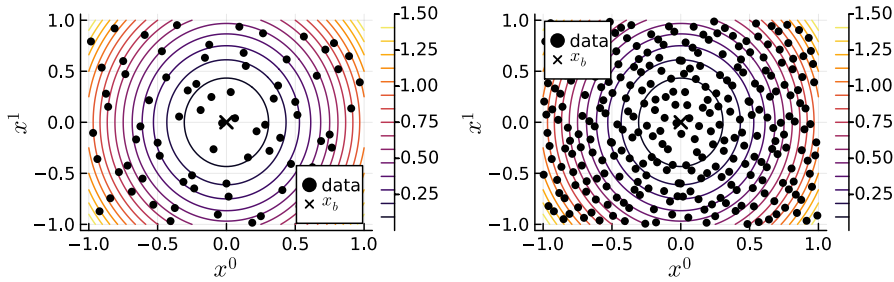


Figure 1: Training data points projected to the  $(x^0, x^1)$ -plane of  $\xi_{80}$  (left) and  $\xi_{300}$  (right).

Here

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

## 5. Numerical experiments

### 5.1. Continuous Lagrangians

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

$$L_{\text{ref}}(x, \dot{x}) = \frac{1}{2} \|\dot{x}\|^2 - \frac{1}{2} \|x\|^2 + \alpha x^0 x^1, \quad x = (x^0, x^1) \in \mathbb{R}^2, (x, \dot{x}) \in T\mathbb{R}^2 \quad (25)$$

with coupling constant  $\alpha = 0.1$ . Here  $\bar{x}^{(j)} = (x^{(j)}, \dot{x}^{(j)})$ ,  $j = 1, \dots, M$  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  $K(\bar{x}, \bar{y}) = \exp(-\frac{1}{2}(\bar{x} - \bar{y})^2)$  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,  $\text{Acc}_{\bar{x}}(L_M)$  refers to the solution of  $\text{EL}(L_M)(\bar{x}, \ddot{x}) = 0$  for  $\ddot{x} \in \mathbb{R}^2$ .

Figure 1 displays the location of training data in  $\Omega$  projected to the  $(x^0, x^1)$ -plane. Figure 2 compares the variances of  $\text{EL}_{\hat{x}}(\xi_M)$  for  $M = 80, 300$  for points of the form  $\hat{x} = (\bar{x}, \ddot{x})$  with  $\bar{x} = (x^0, x^1, 0, 0) \in \Omega$  and  $\bar{x} = (x^0, 0, \dot{x}^0, 0) \in \Omega$  with  $\ddot{x} = \text{Acc}_{\bar{x}}(L_M)$ . One observes that the variance decreases as more data points are used. This experiment 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.

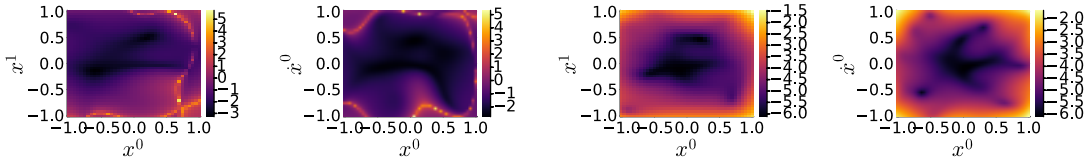


Figure 2: Plots of variances  $\log_{10}(\|\text{var}(\text{EL}(\xi_M))\|)$  for  $M = 80$  (two left plots) and  $M = 300$  (two right plots) over  $(x^0, x^1, 0, 0)$ -plane and  $(x^0, 0, x^0, 0)$ -plane. (Ranges of colourbars vary.)

Figure 3 shows a motion computed by solving<sup>2</sup>  $\text{EL}(L_M) = 0$  with initial data  $\bar{x} = (0.2, 0.1, 0, 0)$  on the time interval  $[0, 100]$ . In the plots of the first row, colours indicate the norm of the variance of  $\text{EL}(\xi_M)$  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  $H_M = \text{Ham}(L_M)$  as well as  $H_M \pm 0.2\sigma_{H_M}$ . Here  $\sigma_{H_M}$  denotes the standard deviation  $\sqrt{\text{var}\text{Ham}(\xi_M)}$ . 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  $\bar{x} = (x^0, x^1, 0, 0) \in \Omega$  and  $\bar{x} = (x^0, 0, x^0, 0) \in \Omega$ . As the magnitudes of errors vary widely,  $\log_{10}$  is applied before plotting, i.e. we show the quantity

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

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  $\text{err}_{\text{Acc}}$ , i.e. of

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

The data for the plot in Figure 6 was computed for the 1d harmonic oscillator  $L_{\text{ref}}(x) = \frac{1}{2}\dot{x}^2 - \frac{1}{2}x^2$  with  $(x, \dot{x}) \in [-1, 1]^2$  in quadruple precision. For each  $M \in \{2^1, 2^2, \dots, 2^6\}$  the error  $\text{err}_{\text{Acc}}(\bar{x})$  was evaluated on a uniform mesh with  $10 \times 11$  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  $10^{-11}$ .

<sup>2</sup>Computations were performed using DifferentialEquations.jl[33]. Comparison with a trajectory computed using the variational midpoint rule [23] (step-size  $h = 0.01$ ) shows a maximal difference in the  $x$ -component smaller than  $3.5 \times 10^{-4}$  ( $M = 300$ ) along the trajectory.



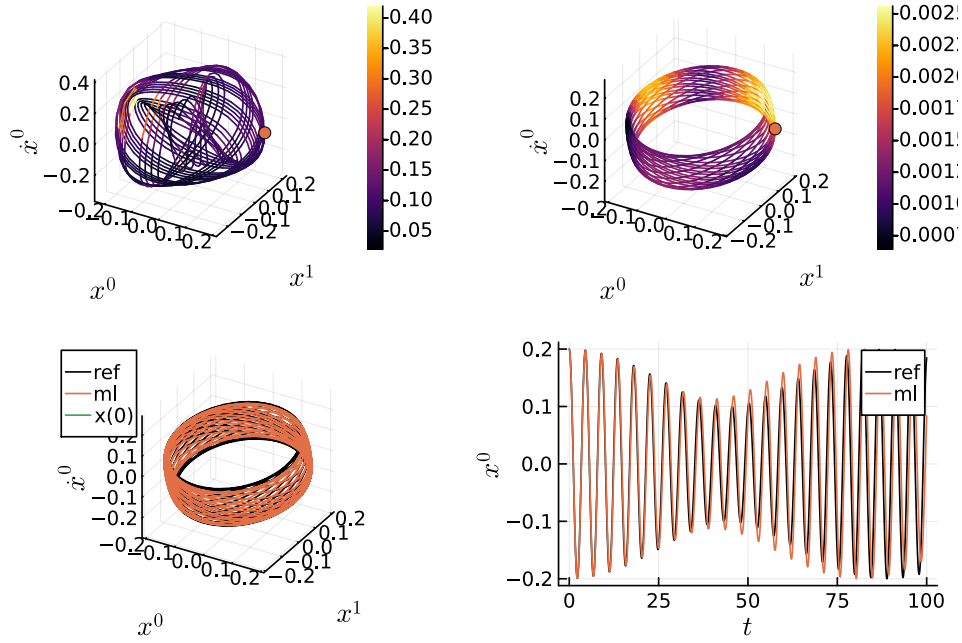


Figure 3: Top row: motion of  $\xi_{80}$  (left) and  $\xi_{300}$  (right) with variance  $\|\text{var}(\text{EL}(\xi_M))\|$  encoded as colours (ranges of colourbars vary). Bottom row: motions of  $\xi_{300}$  compared to reference.

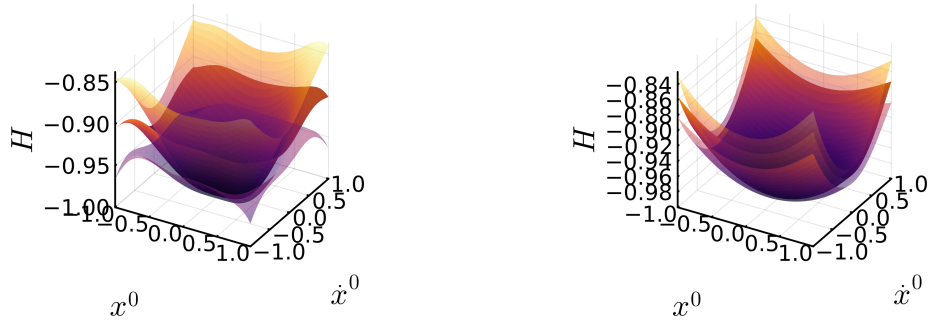


Figure 4: Mean of Hamiltonian  $\text{Ham}(\xi_{80})$ ,  $\text{Ham}(\xi_{300})$  over  $(x^0, 0, \dot{x}^0, 0)$  plus/minus 20% standard deviation.

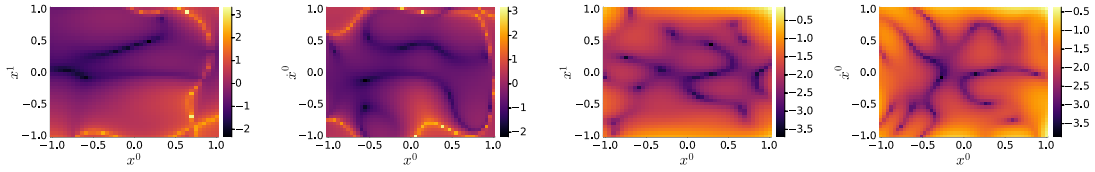


Figure 5:  $\log_{10}$  norm of error of predicted acceleration  $\ddot{x}$  for  $\text{Acc}(\xi_M)$  over  $x^0, x^1$  plane and  $x^0, \dot{x}^0$  plane for  $M = 80$  (left two plots) and  $M = 300$  (right two plots). (The ranges of colourbars vary.)

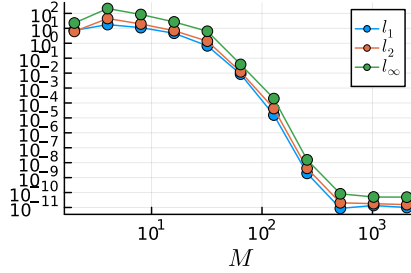


Figure 6: Convergence of  $\text{Acc}(L_M)$  to true acceleration data.

## 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  $j = 1, \dots, M$ . 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  $L_{\text{ref}}$  from  $[0, 3h]$  using the 2nd order accurate variational midpoint rule [23] with step-size  $h_{\text{internal}} = h/10$ . These dynamics are considered as true for the purpose of this experiment. Training data is visualised in Figure 7.

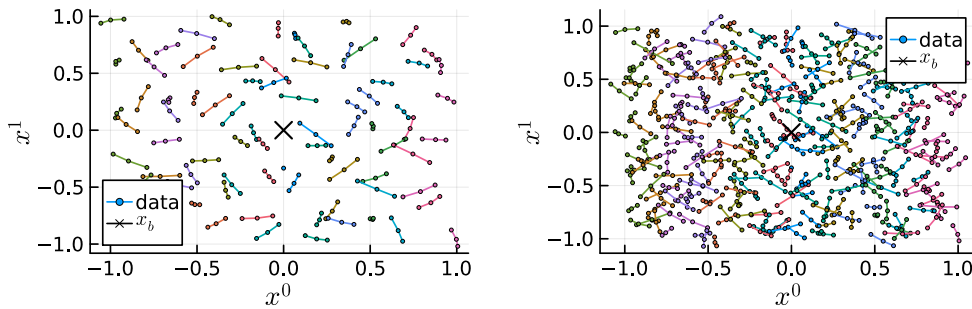


Figure 7: Training data. Each line connects snapshots points that constitute a training data point  $\hat{x}$ . Left:  $M = 80$ , right:  $M = 300$ .

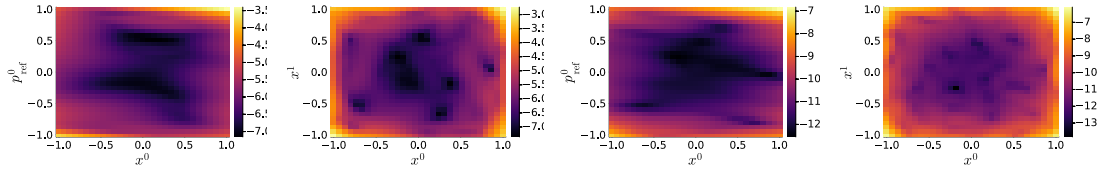


Figure 8: Plots of variances  $\log_{10}(\|\text{var}(\text{EL}(\xi_M))\|)$  for  $M = 80$  (left two plots) and  $M = 300$  (right two plots) over  $(x, p_{\text{ref}}) = (x^0, x^1, 0, 0)$ -plane and  $(x, p_{\text{ref}}) = (x^0, 0, p_{\text{ref}}^0, 0)$ -plane. (Ranges of colourbars vary.)

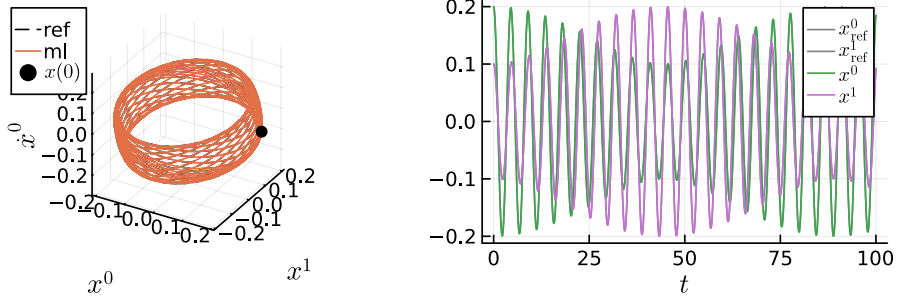


Figure 9: The motion of  $\xi_{300}$  and the true motion are indistinguishable.

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  $L_{\text{ref}}$ . Here  $p$  refers to the conjugate momentum of  $L_{\text{ref}}$ . The plots display heatmaps of  $\log_{10}(\|\text{var}(\text{DEL}_{\hat{x}}(\xi_M))\|)$ .

Figure 9 shows a motion for  $t \in [0, 100]$  of  $\xi_{300}$  with same initial data as in Figure 3. 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](https://github.com/Christian-Offen/Lagrangian_GP).

## 6. Convergence Analysis

This section contains convergence theorems for the considered method for regular continuous Lagrangians (Theorem 1) and discrete Lagrangians (Theorem 2) in the infinite-data limit as observations become topologically dense, i.e. as the maximal distance between data points converges to zero.

## 6.1. Continuous Lagrangians

### 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  $L_{\text{ref}} \in \mathcal{C}^2(\bar{\Omega})$  (definition of  $\mathcal{C}^2(\bar{\Omega})$  below). Assume that  $\{(x^{(j)}, \dot{x}^{(j)})\}_j \subset \Omega$  is topologically dense. Let  $K$  be a 4-times continuously differentiable kernel on  $\Omega$ ,  $\bar{x}_b \in \Omega$ ,  $r_b \in \mathbb{R}$ ,  $p_b \in \mathbb{R}$  and assume that  $L_{\text{ref}}$  is contained in the reproducing kernel Hilbert space  $(U, \|\cdot\|_U)$  to  $K$  and fulfils the normalisation condition*

$$\Phi_N(L_{\text{ref}}) = (p_b, r_b) \quad \text{with} \quad \Phi_N(L) = \left( \frac{\partial L}{\partial \dot{x}}(\bar{x}_b), L(\bar{x}_b) \right). \quad (26)$$

*Assume that  $U$  embeds continuously into  $\mathcal{C}^2(\bar{\Omega})$ . 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  $\xi$  conditioned on the first  $j$  observations and the normalisation conditions*

$$\text{EL}(\xi)(x^{(i)}, \dot{x}^{(i)}, \ddot{x}^{(i)}) = 0 \quad (\forall i \leq j), \quad \Phi_N(\xi) = (p_b, r_b) \quad (27)$$

*converges in  $\|\cdot\|_U$  and in  $\|\cdot\|_{\mathcal{C}^2(\bar{\Omega})}$  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.  $\text{EL}(L_{(\infty)})(\hat{x})$  for all  $\hat{x} = (x, \dot{x}, \ddot{x})$  with  $(x, \dot{x}) \in \Omega$  and  $\text{EL}(L_{\text{ref}})(\hat{x}) = 0$ .*
- *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. However, it be relaxed to twice continuously differentiable as the proof will show. □*

### 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  $\bar{\Omega}$

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

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

$$\|f\|_{\mathcal{C}^m(\bar{\Omega}, \mathbb{R}^k)} = \max_{0 \leq |\alpha| \leq m} \sup_{\bar{x} \in \bar{\Omega}} \|\partial^\alpha f(\bar{x})\|. \quad (28)$$

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

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

$$g_{\text{ref}}(x, \dot{x}) = \left( \frac{\partial^2 L_{\text{ref}}}{\partial \dot{x} \partial \dot{x}}(x, \dot{x}) \right)^{-1} \left( \frac{\partial L_{\text{ref}}}{\partial x}(x, \dot{x}) - \frac{\partial^2 L_{\text{ref}}}{\partial x \partial \dot{x}}(x, \dot{x}) \cdot \dot{x} \right) \quad (29)$$

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

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

$$\begin{aligned} \Phi^{(\infty)}(L)(x, \dot{x}) &= \text{EL}(L)(x, \dot{x}, g_{\text{ref}}(x, \dot{x})) \\ &= \frac{\partial^2 L}{\partial \dot{x} \partial \dot{x}}(x, \dot{x}) \cdot g_{\text{ref}}(x, \dot{x}) + \frac{\partial^2 L}{\partial x \partial \dot{x}}(x, \dot{x}) \cdot \dot{x} - \frac{\partial L}{\partial x}(x, \dot{x}) \end{aligned} \quad (30)$$

is bounded. □

PROOF A direct application of the triangle inequality shows

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

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

$$\begin{aligned} \Phi_j: \mathcal{C}^2(\bar{\Omega}) &\rightarrow \mathbb{R}^d, & \Phi_j(L) &= \Phi^{(\infty)}(L)(\bar{x}^{(j)}) \\ \Phi^{(j)}: \mathcal{C}^2(\bar{\Omega}) &\rightarrow (\mathbb{R}^d)^j, & \Phi^{(j)} &= (\Phi_1, \dots, \Phi_j). \end{aligned}$$

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

$$\Phi_N: \mathcal{C}^2(\bar{\Omega}) \rightarrow \mathbb{R}^{d+1}, \quad \Phi_N(L) = \left( \frac{\partial L}{\partial \dot{x}}(\bar{x}_b), L(\bar{x}_b) \right), \quad (31)$$

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

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

**Assumption 1** Assume that there is a Hilbert space  $U$  with continuous embedding  $U \hookrightarrow \mathcal{C}^2(\overline{\Omega})$  such that

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

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

The affine linear subspace

$$\begin{aligned} A^{(j)} &= \{L \in U \mid \Phi_b^{(j)}(L) = y^{(j)}\} \quad (j \in \mathbb{N}) \\ A^{(\infty)} &= \{L \in U \mid \Phi_b^{(\infty)}(L) = y^{(\infty)}\} \end{aligned}$$

are closed and non empty in  $U$  by Assumption 1 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:

$$\begin{aligned} L_{(j)} &= \arg \min_{L \in A^{(j)}} \|L\|_U \\ L_{(\infty)} &= \arg \min_{L \in A^{(\infty)}} \|L\|_U \end{aligned} \tag{32}$$

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

**Remark 8** In the machine learning setting,  $U$  is the reproducing kernel Hilbert space related to a kernel  $K: \Omega \times \Omega \rightarrow \mathbb{R}$ . Assume the domain of  $\Omega$  is locally Lipschitz. When  $K$  is the squared exponential kernel, for instance, its reproducing kernel Hilbert space embeds into any Sobolev space  $W^m(\Omega)$  ( $m > 1$ ) [8, 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 (32) coincides with the conditional mean of the Gaussian process  $\xi$  conditioned on  $\Phi_b^{(j)}(\xi) = y^{(j)}$ .  $\square$

**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  $A^{(1)} \supseteq A^{(2)} \supseteq A^{(3)} \supseteq \dots$  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  $L_{(\infty)}^\dagger \in U$ . (This follows from the Banach-Alaoglu theorem and the Eberlein-Šmulian theorem [12].) By the weak lower semi-continuity of the norm, we obtain

$$\|L_{(\infty)}^\dagger\|_U \leq \liminf_{i \rightarrow \infty} \|L_{(j_i)}\|_U \leq \|L_{(\infty)}\|_U. \tag{33}$$

**Lemma 4** The weak limit  $L_{(\infty)}^\dagger$  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  $L_{(\infty)}^\dagger \in A^{(\infty)}$ , we have  $\|L_{(\infty)}\|_U \leq \|L_{(\infty)}^\dagger\|_U$  since  $L_{(\infty)}$  is the global minimiser of the minimisation problem of (32). Together with (33) we conclude  $\|L_{(\infty)}^\dagger\|_U = \|L_{(\infty)}\|_U$  and, by the uniqueness of the minimiser  $L_{(\infty)}$ , the equality  $L_{(\infty)}^\dagger = L_{(\infty)}$ . 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

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

such that  $\lim_{i \rightarrow \infty} \|L_{(j_i)}\|_U = \|L_{(\infty)}\|_U$ . Together with  $L_{(j_i)} \rightharpoonup L_{(\infty)}$  we conclude strong convergence  $L_{(j_i)} \rightarrow L_{(\infty)}$  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  $\bar{x} \in \Omega$ . As the sequence  $\Omega_0 = (\bar{x}^{(m)})_{m=1}^\infty$  is dense in  $\Omega$ , there exists a subsequence  $(\bar{x}^{(m_l)})_{l=1}^\infty$  converging to  $\bar{x}$ . We have

$$\Phi_b^{(\infty)}(L_{(\infty)}^\dagger)(\bar{x}) = \lim_{l \rightarrow \infty} \Phi_b^{(\infty)}(L_{(\infty)}^\dagger)(\bar{x}^{(m_l)}) \quad (34)$$

$$= \lim_{l \rightarrow \infty} \underbrace{\lim_{i \rightarrow \infty} \Phi_b^{(\infty)}(L_{(j_i)})(\bar{x}^{(m_l)})}_{\stackrel{(*)}{=} 0} = 0. \quad (35)$$

For this, in (34) we use that  $\Phi_b^{(\infty)}(L_{(\infty)}^\dagger) \in \mathcal{C}^0(\bar{\Omega})$ . Equality in (35) follows because each projection to a component of  $\Phi_b^{(\infty)}(\cdot)(\bar{x}^{(m_l)}): U \rightarrow \mathbb{R}^d \times \mathbb{R}^{d+1}$  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  $\Phi_b^{(\infty)}(L_{(j_i)})(\bar{x}^{(m_l)}) = 0$ .

From  $\Phi_b^{(\infty)}(L_{(\infty)}^\dagger)(\bar{x}) = 0$  for all  $\bar{x} \in \Omega$  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) As by general theory (see Remark 4 or [30, Thm 12.5]), the conditional mean  $L_{(j)}$  is the unique minimiser of the problem (32). Theorem 1 is, therefore, a direct consequence of Proposition 5. ■

## 6.2. Discrete Lagrangian

### 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  $g: \Omega_a \rightarrow \Omega_b$  related to a discrete Lagrangian  $L_d^{\text{ref}} \in \mathcal{C}^1(\bar{\Omega})$ , i.e.*

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

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

$$\Phi_N(L_d^{\text{ref}}) = (p_b, r_b) \quad \text{with} \quad \Phi_N(L_d) = (-\nabla_2 L_d(\Gamma_b), L_d(\Gamma_b)) \quad (36)$$

*and that  $U$  embeds continuously into  $\mathcal{C}^1(\bar{\Omega})$ . 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  $\xi$  conditioned on the first  $j$  observations and the normalisation conditions*

$$\text{DEL}(\xi)(\hat{x}^{(i)}) = 0 \quad (\forall i \leq j), \quad \Phi_N(\xi) = (p_b, r_b) \quad (37)$$

*converges in  $\|\cdot\|_U$  and in  $\|\cdot\|_{\mathcal{C}^1(\bar{\Omega})}$  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.  $\text{DEL}(L_{d,(\infty)})(\hat{x}) = 0$  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  $\text{DEL}(L_d^{\text{ref}})(\hat{x}) = 0$ .*
- *Moreover,  $L_d$  is the unique minimizer of  $\|\cdot\|_U$  among all discrete Lagrangians in  $U$  with the properties above.  $\square$*

**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.  $\square$*



### 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  $\hat{\Omega} = \{(x_0, x_1, x_2) \mid (x_0, x_1) \in \Omega_a, (x_1, x_2) \in \Omega_b\}$  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  $\Omega_a$ .

**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  $L_d^{\text{ref}}: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  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  $\Omega_a$  and  $\Omega_b$  that are neighbourhoods of compact sections of the diagonal in  $\mathbb{R}^d \times \mathbb{R}^d$ .  $\square$

We consider the discrete Lagrangian operator

$$\begin{aligned} \text{DEL}: \mathcal{C}^1(\bar{\Omega}) &\rightarrow \mathcal{C}^0(\bar{\Omega}, \mathbb{R}^d) \\ \text{DEL}(L_d)(x_0, x_1, x_2) &= \nabla_2 L_d(x_0, x_1) + \nabla_1 L_d(x_1, x_2). \end{aligned} \quad (38)$$

Here  $\nabla_j L_d$  denotes the partial derivatives with respect to the  $j$ th 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  $L_d^{\text{ref}} \in \mathcal{C}^1(\bar{\Omega})$  with globally Lipschitz continuous flow map  $g: \Omega_a \rightarrow \Omega_b$ , i.e.  $\text{DEL}(L_d^{\text{ref}})(x_0, g(x_0, x_1)) = 0$  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  $\Phi^{(\infty)}: \mathcal{C}^1(\bar{\Omega}) \rightarrow \mathcal{C}^0(\bar{\Omega}_a, \mathbb{R}^d)$  with*

$$\Phi^{(\infty)}(L_d)(x_0, x_1, x_2) = \text{DEL}(L_d)(x_0, g(x_0, x_1)) \quad (39)$$

*is bounded.*  $\square$

**PROOF** Indeed,  $g$  extends to a globally Lipschitz continuous map  $g: \bar{\Omega}_a \rightarrow \bar{\Omega}_b$  such that  $\Phi^{(\infty)}: \mathcal{C}^1(\bar{\Omega}) \rightarrow \mathcal{C}^0(\bar{\Omega}_a, \mathbb{R}^d)$  is a well-defined map between Banach spaces defined via (39). Let  $\|L_d\|_{\mathcal{C}^1(\bar{\Omega})} \leq 1$ . In particular,

$$\sup_{(x_0, x_1) \in \Omega_a} \|\nabla_2 L_d(x_0, x_1)\| \leq 1 \quad \text{and} \quad \sup_{(x_1, x_2) \in \Omega_b} \|\nabla_1 L_d(x_1, x_2)\| \leq 1. \quad (40)$$

Therefore, by the triangle inequality

$$\begin{aligned} \sup_{(x_0, x_1) \in \Omega_a} \text{DEL}(L_d)(x_0, g(x_0, x_1)) &\leq 1 + \sup_{(x_0, x_1) \in \Omega_a} \|\nabla_2 L_d(g(x_0, x_1))\| \\ &\leq 1 + \sup_{(x_1, x_2) \in \Omega_b} \|\nabla_1 L_d(x_1, x_2)\| \leq 2. \end{aligned} \quad (41) \quad \blacksquare$$

We can now proceed in direct analogy to the continuous setting (Section 6.1.2) with  $L$  replaced by  $L_d$  and the functional  $\Phi_N$  of (31) (normalisation conditions) replaced by the corresponding functional for discrete Lagrangians. The details are provided in the following.

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

$$\begin{aligned}\Phi_j: \mathcal{C}^1(\bar{\Omega}) &\rightarrow \mathbb{R}^d, & \Phi_j(L_d) &= \Phi^{(\infty)}(L_d)(\bar{x}^{(j)}) \\ \Phi^{(j)}: \mathcal{C}^1(\bar{\Omega}) &\rightarrow (\mathbb{R}^d)^j, & \Phi^{(j)} &= (\Phi_1, \dots, \Phi_j).\end{aligned}$$

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

$$\Phi_N: \mathcal{C}^1(\bar{\Omega}) \rightarrow \mathbb{R}^{d+1}, \quad \Phi_N(L) = (-\nabla_1 L_d(\bar{x}_b), L_d(\bar{x}_b)), \quad (42)$$

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

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

In analogy to Assumption 1 we consider the following assumption.

**Assumption 2** *Assume that there is a Hilbert space  $U$  with continuous embedding  $U \hookrightarrow \mathcal{C}^1(\bar{\Omega})$  such that*

$$\{L_d \in \mathcal{C}^1(\bar{\Omega}) \mid \Phi_b^{(\infty)}(L_d) = y^{(\infty)}\} \cap U \neq \emptyset$$

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

The affine linear subspace

$$\begin{aligned}A^{(j)} &= \{L_d \in U \mid \Phi_b^{(j)}(L_d) = y^{(j)}\} \quad (j \in \mathbb{N}) \\ A^{(\infty)} &= \{L_d \in U \mid \Phi_b^{(\infty)}(L_d) = y^{(\infty)}\}\end{aligned}$$

are closed in  $U$  and not empty by Assumption 2. Therefore, the following extremisation problems constitute convex optimisation problems on  $U$  with unique minima in  $A^{(j)}$  or  $A^{(\infty)}$ , respectively:

$$\begin{aligned}L_{d(j)} &= \arg \min_{L_d \in A^{(j)}} \|L_d\|_U \\ L_{d(\infty)} &= \arg \min_{L_d \in A^{(\infty)}} \|L_d\|_U\end{aligned} \quad (43)$$

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  $\|\cdot\|_{\mathcal{C}^1(\bar{\Omega})}$ .  $\square$*

PROOF The proof is in complete analogy to Proposition 5.  $\blacksquare$

PROOF (THEOREM 2) The unique minimisers  $L_{d(j)}$  in (43) are the conditional means considered in Theorem 2 as by Remark 4. Thus, Theorem 2 follows from Proposition 6.  $\blacksquare$

## 7. Summary

We have introduced a method to learn general continuous Lagrangians and discrete Lagrangians from observational data of dynamical system that are governed by variational ordinary differential equations. The method is based on kernel-based, meshless collocation methods for solving partial differential equations [36]. 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.

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 normalisation 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 [30]. 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 future research we will extend the method to dynamical systems governed by variational partial differential equations. Moreover, it is of interest to identify and prove convergence rates of the proposed method. A further direction is the combination with detection methods for Lie group variational symmetries [11, 20] or with detection methods of travelling waves [26, 28]. 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](https://github.com/Christian-Offen/Lagrangian_GP). An archived version [25] of release v1.0 of the GitHub repository is openly available at <https://doi.org/10.5281/zenodo.11093645>.

## Appendices

### A. 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 [28].

**Proposition 7** *Let  $\bar{x}_b = (x_b, \dot{x}_b) \in T\mathbb{R}^d \cong \mathbb{R}^d \times \mathbb{R}^d$  and  $\dot{L}$  a Lagrangian with  $\frac{\partial \dot{L}}{\partial \dot{x} \partial \dot{x}}(\bar{x}_b)$  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  $\dot{L}$  and*

$$L(\bar{x}_b) = c_b, \quad \text{Mm}(L)(\bar{x}_b) = \frac{\partial L}{\partial \dot{x}}(\bar{x}_b) = p_b, \quad N_\omega(L)(\bar{x}_b) = \left| \det \left( \frac{\partial^2 L}{\partial \dot{x} \partial \dot{x}}(\bar{x}_b) \right) \right| = c_\omega. \quad (44)$$

□

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

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

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

The condition  $N_\omega(L)(\bar{x}_b) = c_\omega > 0$  may be compared to the regularisation strategies for training Lagrangians modelled as neural networks in [28]: denoting observation data by  $\hat{x}^{(j)} = (x^{(j)}, \dot{x}^{(j)}, \ddot{x}^{(j)})$ , in [28] (transferred to our continuous ode setting) parametrises  $L$  as a neural network and considers the minimisation of a loss function  $\ell = \ell_{\text{data}} + \ell_{\text{reg}}$  with data consistency term

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

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

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

The corresponding statement for discrete Lagrangians is as follows.

**Proposition 8** Let  $\bar{x}_b = (x_{0b}, x_{1b}) \in \mathbb{R}^d \times \mathbb{R}^d$  and  $\mathring{L}_d$  a discrete Lagrangian with  $\text{Mm}^-(\bar{x}_b)$  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

$$L_d(\bar{x}_b) = c_b, \quad \text{Mm}^-(L_d)(\bar{x}_b) = p_b, \quad N_\omega^-(L_d)(\bar{x}_b) = \left| \det \left( \frac{\partial^2 L_d}{\partial x_0 \partial x_1}(\bar{x}_b) \right) \right| = c_\omega. \quad (45)$$

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

$$\rho = \sqrt[2]{\frac{c_\omega}{\mathring{c}_\omega}}, \quad F(x) = x^\top (p_b - \rho \mathring{p}_b), \quad c = c_b - \rho \mathring{c}_b - (x_{1b} - x_{0b})^\top (p_b - \rho \mathring{p}_b).$$

Now the Lagrangian  $L_d = \rho \mathring{L}_d + \Delta_t F + c$  is equivalent to  $L_d$  and fulfils (45).  $\blacksquare$

Again, the condition  $N_\omega^-(L)(\bar{x}_b) = c_\omega > 0$  may be compared to the regularisation strategies for training discrete Lagrangians modelled as neural networks in [28]: denoting observation data by  $\hat{x}^{(j)} = (x_0^{(j)}, x_1^{(j)}, x_2^{(j)})$ , in [28] (when transferred to our discrete ode setting) parametrises  $L_d$  as a neural network and considers the minimisation of a loss function  $\ell = \ell_{\text{data}} + \ell_{\text{reg}}$  with data consistency term

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

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

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

## B. 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  $\Phi^\pm: \mathbb{R}^d \times \mathbb{R}^d \rightarrow T^*\mathbb{R}^d$  [23] with

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

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

$$\begin{aligned}
\text{Sympl}^-(L_d) &= \sum_{s=1}^d dx_0^s \wedge d \left( -\frac{\partial L_d}{\partial x_0^s} \right) = \sum_{r,s=1}^d -\frac{\partial^2 L_d}{\partial x_0^s \partial x_0^r} dx_0^s \wedge dx_0^r - \frac{\partial^2 L_d}{\partial x_0^s \partial x_1^r} dx_0^s \wedge dx_1^r \\
&= \sum_{r,s=1}^d -\frac{\partial^2 L_d}{\partial x_0^s \partial x_1^r} dx_0^s \wedge dx_1^r \\
\text{Sympl}^+(L_d) &= \sum_{s=1}^d dx_1^s \wedge d \left( \frac{\partial L_d}{\partial x_1^s} \right) = \sum_{r,s=1}^d \frac{\partial^2 L_d}{\partial x_1^s \partial x_0^r} dx_1^s \wedge dx_0^r + \frac{\partial^2 L_d}{\partial x_1^s \partial x_1^r} dx_1^s \wedge dx_1^r \\
&= \sum_{r,s=1}^d \frac{\partial^2 L_d}{\partial x_1^s \partial x_0^r} dx_1^s \wedge dx_0^r
\end{aligned}$$

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

The 2-form corresponds to the notion of a *discrete Lagrangian symplectic form* in [23, §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] 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.
- [4] J F Carinena and L 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.
- [5] 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:arXiv:2103.05632.
- [6] 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.

- [7] 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>.
- [8] 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.
- [9] Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks, 2020. doi:10.48550/ARXIV.2003.04630.
- [10] 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.
- [11] 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.
- [12] Joseph Diestel. *Sequences and Series in Banach Spaces*. Springer New York, 1984. doi:10.1007/978-1-4612-5200-9.
- [13] 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.
- [14] 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.
- [15] I.M. Gelfand, S.V. Fomin, and R.A. Silverman. *Calculus of Variations*. Dover Books on Mathematics. Dover Publications, 2000.
- [16] 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.
- [17] 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.
- [18] Jianyu Hu, Juan-Pablo Ortega, and Daiying Yin. A structure-preserving kernel method for learning Hamiltonian systems, 2024. arXiv:2403.10070.

- [19] 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.
- [20] 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.
- [21] G. Marmo and G. 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.
- [22] Giuseppe Marmo and C. 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.
- [23] Jerrold E. Marsden and Matthew West. Discrete mechanics and variational integrators. *Acta Numerica*, 10:357–514, 2001. doi:10.1017/S096249290100006X.
- [24] 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.
- [25] Christian Offen. Software: Christian-Offen/Lagrangian\_GP: Initial release of GitHub Repository, 4 2024. doi:10.5281/zenodo.11093645.
- [26] 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.
- [27] Christian Offen and Sina Ober-Blöbaum. Symplectic integration of learned Hamiltonian systems. *Chaos*, 32(1):013122, 1 2022. doi:10.1063/5.0065913.
- [28] 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.
- [29] Juan-Pablo Ortega and Daiying Yin. Learnability of linear port-Hamiltonian systems, 2023. arXiv:2303.15779.
- [30] 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.



- [31] Hong Qin. Machine learning and serving of discrete field theories. *Scientific Reports*, 10(1), 11 2020. doi:10.1038/s41598-020-76301-0.
- [32] 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>.
- [33] 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.
- [34] 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.
- [35] Tomáš Roubíček. *Calculus of Variations*, pages 1–38. John Wiley & Sons, Ltd, 2015. doi:10.1002/3527600434.eap735.
- [36] Robert Schaback and Holger Wendland. Kernel techniques: From machine learning to meshless methods. *Acta Numerica*, 15:543–639, 2006. doi:10.1017/S0962492906270016.
- [37] 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.
- [38] Mats Vermeeren. Modified equations for variational integrators. *Numerische Mathematik*, 137(4):1001–1037, 6 2017. doi:10.1007/s00211-017-0896-4.