Physics-enhanced Gaussian Process Variational Autoencoder

Thomas Beckers THOMAS.BECKERS@VANDERBILT.EDU
Department of Computer Science, Vanderbilt University, Nashville, USA

Qirui Wu WUQIRUI@SEAS.UPENN.EDU
George J. Pappas PAPPASG@SEAS.UPENN.EDU
Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, USA

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

Variational autoencoders allow to learn a lower-dimensional latent space based on high-dimensional input/output data. Using video clips as input data, the encoder may be used to describe the movement of an object in the video without ground truth data (unsupervised learning). Even though the object's dynamics is typically based on first principles, this prior knowledge is mostly ignored in the existing literature. Thus, we propose a physics-enhanced variational autoencoder that places a physical-enhanced Gaussian process prior on the latent dynamics to improve the efficiency of the variational autoencoder and to allow physically correct predictions. The physical prior knowledge expressed as linear dynamical system is here reflected by the Green's function and included in the kernel function of the Gaussian process. The benefits of the proposed approach are highlighted in a simulation with an oscillating particle.

Keywords: physics-enhance learning, scientific machine learning, variational autoencoders, Gaussian processes

1. Introduction

Variational autoencoders (VAEs) have been one of the most popular approaches to unsupervised learning of complex distributions (Doersch, 2016). Their effectiveness has been proven in several examples, such as for handwritten digits (Kingma and Welling, 2013), faces (Rezende et al., 2014), CIFAR images (Gregor et al., 2015), segmentation (Sohn et al., 2015), and prediction of the future from static images (Walker et al., 2016). Further, VAE can not only be used to learn the latent state for static objects but also for time-transient inputs such as videos. In this case, there exists a latent time series to describe the evolution of the latent state over time. VAEs for videos have been used in the context of anomalies detection (Waseem et al., 2022), long-horizon predictions (Saxena et al., 2021), learning spatial knowledge for mobile robots (Nagano et al., 2022), and training data generation for autonomous driving (Amini et al., 2018). In all of these applications, the observed objects are typically subject to certain physical rules as they exist and operate in real world environments. However, this prior knowledge is mostly neglected, which might lead to unrealistic predictions and data-hungry algorithms.

In this article, we consider the learning of a physical grounded latent time series of a video showing a moving object. The object's dynamics is based on physical laws encoded as linear latent dynamics, which can be excited by an external, unknown input (see Figure 1). A simple example is a mass-spring-damper system with an external excitation generated by an electromechanical actuator.

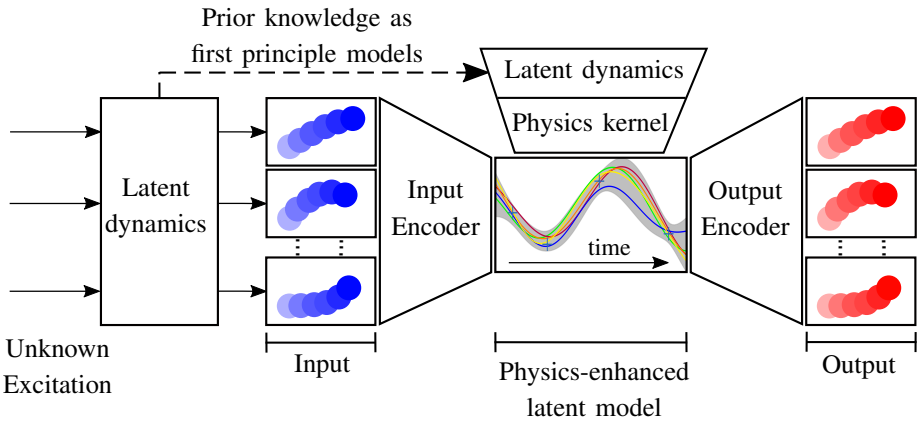

Figure 1: Model of the physics-enhanced Gaussian process variational autoencoder (PEGP-VAE). The input consists of a batch of video sequences of a physical system generated by latent dynamics with unknown excitations, e.g., the position of a ball over time. The latent dynamics are used to formulate a physics-enhanced kernel function to model the latent time series with a Gaussian process.

Other examples include pedestrian movements, where the pedestrians are modeled as masses driven by an external force, or micro-particles in electromagnetic fields.

Although learning approaches such as neural networks are highly flexible in describing latent time series, physical knowledge expressed as differential equations is much less restricted by data availability, as they can make accurate predictions even without training data (Hou and Wang, 2013). Therefore, we aim to combine the best of both worlds: Using physical prior knowledge for the latent space with expressive models for the unknown external excitation. For this purpose, we leverage Gaussian processes with prior knowledge expressed as linear differential equation as prior for the latent time series. Gaussian processes (GPs) have been developed as powerful function regressors. A GP connects every point in a continuous input space with a normally distributed random variable such that any finite group of those infinitely many random variables follows a multivariate Gaussian distribution (Rasmussen and Williams, 2006). In contrast to most of the other techniques, GP modeling provides not only a mean function but also a measure for the uncertainty of the prediction.

Contribution: In this article, we propose a physics-enhanced Gaussian process variational autoencoder (PEGP-VAE) bringing together physical prior knowledge encoded as a linear system with a GP prior on the latent dynamics. For this purpose, we use the Green's function of the linear system to construct a linear operator that is included in the kernel function of the GP. The PEGP-VAE is trained with a batch of video sequences consisting of a moving object following the linear dynamics with unknown excitation. Then, new video sequences can be generated with uncertainty quantification based on the posterior variance of the GP. The physical model allows the VAE to be more efficient in training and to make predictions which respect the physical prior.

The remainder of the paper is structured as follows. After introducing the problem statement in Section 1.1, we briefly summarize the background techniques in Section 2, followed by presenting the PEGP-VAE in Section 3. Finally, a simulation is performed in Section 4.

Notation: Vectors and vector-valued functions are denoted with bold characters $\boldsymbol{v}$. The notation $[\boldsymbol{a};\boldsymbol{b}]$ is used for $[\boldsymbol{a}^\top, \boldsymbol{b}^\top]^\top$ and $\boldsymbol{x}^{(1:n)}$ denotes $[\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_n^\top]^\top$. Capital letters A describes matrices. The matrix I is the identity matrix of appropriate dimension. The expression $\mathcal{N}(\mu, \Sigma)$ describes a normal distribution with mean μ and covariance Σ. N_+ and $\mathbb{R}_+$ denote the positive natural and positive real numbers, respectively.

1.1. Problem description

We consider the problem of learning a lower-dimensional, physics-enhanced latent time series based on a batch of video sequences of a moving object. Movement of the object is generated by linear dynamics based on first principles with an unknown (nonlinear) external, time-dependent excitation $\boldsymbol{u}\colon \mathbb{R}_+ \to \mathbb{R}^m$. The latent dynamics is defined by

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x} + B\boldsymbol{u}(t), \quad \boldsymbol{y} = C\boldsymbol{x} \tag{1}$$

with state $\boldsymbol{x} \in \mathbb{R}^n$, output $\boldsymbol{y} \in \mathbb{R}^p$, system matrix $A \in \mathbb{R}^{n\times n}$, input matrix $B \in \mathbb{R}^{n\times m}$, output matrix $C \in \mathbb{R}^{p\times n}$, and time $t \in \mathbb{R}_+$. The matrices A, B, C are assumed to be known except for a finite number n_φ of unknown parameters bundled in a vector $\boldsymbol{\varphi} \in \mathbb{R}^{n_\varphi}$. We consider the existence of n_v video clips in which each clip consists of n_f black-white frames described by $\boldsymbol{v}^{(i)} = [0,1]^{d^2}, i \in \{1, \dots, n_f\}$ with $d \in \mathbb{N}$ pixels in width and height. The frame $\boldsymbol{v}^{(i)}$ is recorded at time $t_i \in \mathbb{R}_+$ with equidistant $t_1, \dots, t_{n_f}$. The goal is to find a latent time series $[\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(n_f)}]$ that describes the evolution of the object over time based on the n_v video clips. The evolution of $\boldsymbol{y}$ shall be consistent with the prior knowledge expressed by the linear system (1). In the remainder of the paper, we will refer to $\boldsymbol{y}$ as latent state that should not be confused with the state $\boldsymbol{x}$ in (1). Note that we do not consider the existence of a ground truth for the latent state $\boldsymbol{y}$.

1.2. Related work

Finding interpretable low-dimensional dynamics from pixels has been considered by exploiting state-space models, e.g., in Fraccaro et al. (2017); Lin et al. (2018); Pearce et al. (2018), which assume an underlying Markov structure to enforce interpretability on latent representations. One of the first papers where GPs are connected with variational autoencoders has been published by Casale et al. (2018). The proposed method is based on a fully factorized approximate posterior that, however, performs poorly in time series and spatial settings (Barber et al., 2011). Fortuin et al. (2020) consider the use of a Gaussian approximate posterior with a tridiagonal precision matrix parameterized by an inference network. Whilst this permits computational efficiency, the parameterization neglects a rigorous treatment of long-term dependencies. Campbell and Liò (2020) has extended this framework to handle more general spatio-temporal data. Finally, Pearce (2020) propose a GP based VAE approach with structured approximate posterior allowing long-term dependencies, and Ashman et al. (2020) generalized this framework to handle missing data. However, these works do not consider using physical prior knowledge in the latent dynamics. Using physics as a prior knowledge in VAEs has been mainly addressed by using neural networks (Luchnikov et al., 2019; Farina et al., 2020; Erichson et al., 2019) which inherently lack information on the uncertainty of the model. Although GPs are highly suitable for the integration of prior knowledge, e.g., in robotics (Geist and Trimpe, 2020; Rath et al., 2021) or more general physical systems (De Bézenac et al., 2019; Hanuka et al., 2020; Wang et al., 2020), the connection of variational autoencoders with physics-enhanced GP priors on the latent time series is still open.

2. Background

2.1. Gaussian Process Models

Let $(\Omega, \mathcal{F}, P)$ be a probability space with the sample space Ω, the corresponding σ-algebra $\mathcal{F}$ and the probability measure P. Consider a vector-valued, unknown time series $\boldsymbol{f}\colon \mathbb{R}_+ \to \mathbb{R}^p$. The measurement $\tilde{\boldsymbol{y}} \in \mathbb{R}^p$ of the series is corrupted by Gaussian noise $\boldsymbol{\eta} \in \mathbb{R}^p$, i.e., $\tilde{\boldsymbol{y}} = \boldsymbol{f}(t) + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \Sigma_n)$ with the positive definite matrix $\Sigma_n = \operatorname{diag}(\sigma_1^2, \dots, \sigma_p^2)$. The function is measured at n_f input values $\{t^{\{j\}}\}_{j=1}^{n_f}$. Together with the resulting measurements $\{\tilde{\boldsymbol{y}}^{\{j\}}\}_{j=1}^{n_f}$, the whole training data set is described by $\mathcal{D} = \{T, Y\}$ with the input training matrix $T = [t^{\{1\}}, t^{\{2\}}, \dots, t^{\{n_f\}}] \in \mathbb{R}^{1\times n_f}$ and the output training matrix $Y = [\tilde{\boldsymbol{y}}^{\{1\}}, \tilde{\boldsymbol{y}}^{\{2\}}, \dots, \tilde{\boldsymbol{y}}^{\{n_f\}}]^\top \in \mathbb{R}^{n_f \times p}$. Now, the objective is to predict the output of the function $\boldsymbol{f}(t^*)$ at a test input $t^* \in \mathbb{R}_+$. The underlying assumption of GP modeling is that the data can be represented as a sample of a multivariate Gaussian distribution using a kernel function k. The joint distribution of the i-th component of $\boldsymbol{f}(t^*)$ is[1]

$$\begin{bmatrix} Y_{:,i} \\ f_i(t^*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(T,T) + \sigma_i^2 I & \boldsymbol{k}(t^*, T) \\ \boldsymbol{k}(t^*, X)^\top & k(t^*, t^*) \end{bmatrix}\right) \tag{2}$$

with the kernel $k\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ as a measure of the correlation of two points (t, t'). The function $K\colon \mathbb{R}^{1\times n_f} \times \mathbb{R}^{1\times n_f} \to \mathbb{R}^{n_f \times n_f}$ is called the Gram matrix $K_{j,l} = k(T_{1,l}, T_{1,j})$ with $j, l \in \{1, \dots, n_f\}$. Each element of the matrix represents the covariance between two elements of the training data T. The vector-valued function $\boldsymbol{k}\colon \mathbb{R}_+ \times \mathbb{R}^{1\times n_f} \to \mathbb{R}^{n_f}$ calculates the covariance between the test input t^* and the input training data T where $k_j = k(t^*, T_{1,j})$ for all $j \in \{1, \dots, n_f\}$. A comparison of the characteristics of the different covariance functions can be found in Bishop and Nasrabadi (2006). The prediction of each component of $\boldsymbol{f}(t^*)$ is derived from the joint distribution (2) and is therefore a Gaussian distributed variable. The conditional probability distribution for the i-th element of the output is defined by the mean and the variance

$$\begin{aligned} \mu_i(\boldsymbol{f}|t^*, \mathcal{D}) =& \boldsymbol{k}(t^*, T)^\top (K + \sigma_i^2 I)^{-1} Y_{:,i} \\ \operatorname{var}_i(\boldsymbol{f}|t^*, \mathcal{D}) =& k(t^*, t^*) - \boldsymbol{k}(t^*, T)^\top (K + \sigma_i^2 I)^{-1} \boldsymbol{k}(t^*, T). \end{aligned} \tag{3}$$

Finally, the q normally distributed components of $\boldsymbol{f}|t^*, \mathcal{D}$ can be combined into a multi-variable Gaussian distribution $\boldsymbol{f}|t^*, \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}(\cdot), \Sigma(\cdot))$ with $\boldsymbol{\mu}(\boldsymbol{f}|t^*, \mathcal{D}) = [\mu(f_1|t^*, \mathcal{D}), \dots, \mu(f_p|t^*, \mathcal{D})]^\top$ and $\Sigma(\boldsymbol{f}|t^*, \mathcal{D}) = \operatorname{diag}(\operatorname{var}(f_1|t^*, \mathcal{D}), \dots, \operatorname{var}(f_p|t^*, \mathcal{D}))$.

2.2. Latent Force Models

In real-world dynamics, physics knowledge, expressed as differential equations, provides useful insight into the mechanism of the system and can be beneficial for understanding and prediction. Alvarez et al. (2013) introduced the latent force model (LFM) that allows incorporating physical prior knowledge into GP models. We consider a LFM with p output functions $y_1, \dots, y_p\colon \mathbb{R}_+ \to \mathbb{R}$ and latent forces $u_1, \dots, u_m\colon \mathbb{R}_+ \to \mathbb{R}$ to define the differential equation

$$\mathcal{L}\,\boldsymbol{y}(t) = \boldsymbol{u}(t), \tag{4}$$

where $\mathcal{L}$ is a linear differential operator (Courant and Hilbert, 2008). Using the latent force model (4), a GP prior is placed on the unknown latent forces $u_i \sim \mathcal{GP}(0, k_{u_i})$. As GPs are closed under linear operators (Rasmussen and Williams, 2006) and $\mathcal{L}$ is linear, each function y_i also defines a GP.

1. For notational convenience, we simplify $K(T, T)$ to K

3. PEGP-VAE

In this section, we propose the physics-enhanced Gaussian process variational auto-encoder, which allows us to integrate physical prior knowledge into the latent dynamics. The goal is to find a lower-dimensional physical representation for the movement of the object in the video clips. As we do not have a ground truth for the latent state, it is an unsupervised learning problem with respect to the latent time series. In the following, we use the notation $\boldsymbol{y}^{(1:n_f)}$ for $[y_1^{(1)}, \dots, y_1^{(n_f)}, \dots, y_p^{(1)}, \dots, y_p^{(n_f)}]^\top$ to describe the state of the object, where $\boldsymbol{y}^{(i)} \in \mathbb{R}^p$ is the latent state at time t_i. In addition, $T = [t_1, \dots, t_{n_f}]$ denotes the vector of the recorded time stamps. For example, the latent state could describe the position of a ball in the video frame as illustrated in Fig. 2. Then, the goal is to find the unknown latent input $\boldsymbol{u}$ such that the evolution of the latent state is consistent with the latent dynamics (1). Thus, we place GP priors on the unknown latent input (excitation) $\boldsymbol{u}$ by $u_i \sim \mathcal{GP}(0, k_{u_i}(T,T))$ for all $i \in \{1, \dots, m\}$. For simplicity, the priors are independent, but extensions to multi-output GPs to model correlations between latent inputs are possible. Then, we model the joint probability distribution $P(\boldsymbol{v}^{(1:n_f)}, \boldsymbol{y}^{(1:n_f)})$ between the video frames $\boldsymbol{v}^{(1:n_f)} \in \mathbb{R}^{n_f d^2}$ and the latent states $\boldsymbol{y}^{(1:n_f)} \in \mathbb{R}^{n_f p}$ by

$$
\begin{aligned}
P(\boldsymbol{v}^{(1:n_f)}, \boldsymbol{y}^{(1:n_f)}) &= \prod_{i=1}^{n_f} P(\boldsymbol{v}^{(i)}|, \boldsymbol{y}^{(i)}) P(\boldsymbol{y}^{(1:n_f)}) \qquad (5) \\
&= \prod_{i=1}^{n_f} \underbrace{\mathcal{B}(\boldsymbol{v}^{(i)}|, p_\theta(\boldsymbol{y}^{(i)}))}_{\text{Pixel model}} \underbrace{\mathcal{N}\left(\boldsymbol{y}^{(1:n_f)} \middle| \boldsymbol{0}, \begin{bmatrix} K_{11}(T,T) & \dots & K_{1p}(T,T) \\ \vdots & & \vdots \\ K_{1p}(T,T)^\top & \dots & K_{pp}(T,T) \end{bmatrix}\right)}_{\text{Latent dynamics}}
\end{aligned}
$$

where $\mathcal{B}(\boldsymbol{v}^{(i)}|, p_\theta(\boldsymbol{y}^{(i)}))$ is a product of d^2 independent Bernoulli distributions over the pixels of the frame $\boldsymbol{v}^{(i)}$ parameterized by a neural network $p_\theta(\boldsymbol{y}^{(i)})$ with parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ similar to Pearce (2020). The latent dynamics is described by a multivariate normal distribution $\mathcal{N}(\cdot|\mu, \Sigma)$ with mean μ and variance Σ as it contains a finite subset of the GP. Next, we show how to include prior knowledge in (5) via a physics-enhanced kernel for Gram matrices $K_{11}, \dots, K_{pp} \in \mathbb{R}^{n_f \times n_f}$.

3.1. Physical Prior Knowledge

Our goal is to encode the latent dynamics (1) in a kernel function. In this way, we use a physics-enhanced GP prior on the latent model of the VAE. Following the idea of Latent force models (Alvarez et al., 2013), we need to find a linear differential operator that describes the time evolution of the latent dynamics (1). In this regard, let $G\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}^{p \times m}$ be the Green's function of the latent dynamics. The Green's function is known to be the impulse response for linear dynamical systems which can be determined by

$$
G(t, t') = C\boldsymbol{e}^{A(t-t')}B \qquad (6)
$$

with the matrix exponential $\boldsymbol{e}$, input matrix B, output matrix C and system matrix A given by (1). The impulse response allows us to compute the solution of the initial-value problem with $\boldsymbol{x}_0 = \boldsymbol{0}$ via convolution

$$
(G * u)(t) = \int_0^t G(t, t')u(t')dt', \qquad (7)
$$

where $*$ denotes the convolution operator. Now, we can build a linear operator as in Section 2.2 using the Green's function and the convolution (7) to create a physics-enhanced kernel. As result, the enhanced kernel $k_{ij}\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$, that describes the covariance between the i-th and j-th dimension of the latent state $\boldsymbol{y}$, is computed by

$$k_{ij}(t,t') = \int_0^t \int_0^{t'} G_{i,:}(t,\tau) \begin{bmatrix} k_{u_1}(\tau,\tau') & & 0 \\ & \ddots & \\ 0 & & k_{u_m}(\tau,\tau') \end{bmatrix} G_{j,:}(t',\tau')^\top d\tau d\tau' \tag{8}$$

for all $i,j \in \{1,\dots,p\}$. Then, the Gram matrices $K_{ij} \in \mathbb{R}^{n_f \times n_f}$ are constructed as stated in Section 2.1.

Remark 1 *For some kernels that are used in the GP prior on the independent, unknown inputs $u_1,\dots,u_m$, there exists an analytic solution for (7). For example, the commonly used squared exponential kernel leads to a closed-form solution (Alvarez et al., 2013).*

Remark 2 *We only need to consider $\boldsymbol{x}_0 = \mathbf{0}$ as initial value as the encoder network can always perform a linear transformation in the case of $\boldsymbol{x}_0 \neq \mathbf{0}$.*

For more detailed information on convolution for kernel functions and the analytical solution for the squared exponential kernel, we refer to Van der Wilk et al. (2017).

3.2. Prediction

Equipped with the physics-enhanced kernel, the goal is to compute the conditional distribution $P(\boldsymbol{y}^{(1:n_f)}|\boldsymbol{v}^{(1:n_f)})$ given the latent states based on a video sequence. For simplicity of notation, we assume that the latent states and the video sequence have the same number of time steps that, however, can be easily adapted.

Due to the Bernoulli distribution term $\mathcal{B}\big(\boldsymbol{v}^{(i)}|, p_\theta(\boldsymbol{y}^{(i)})\big)$, there exists no analytic solution for the posterior. Inspired by Pearce (2020), we propose the following variational approximation

$$\begin{aligned} q(\boldsymbol{y}^{(1:n_f)}|\boldsymbol{v}^{(1:n_f)}) &= \frac{1}{L(\boldsymbol{v}^{(1:n_f)})} \prod_{i=1}^{n_f} q^*_\Phi(\boldsymbol{y}^{(i)}|\boldsymbol{v}^{(i)}) P(\boldsymbol{y}^{(1:n_f)}) \qquad (9) \\ &= \underbrace{\prod_{i=1}^{n_f} \mathcal{N}\left(\boldsymbol{y}^{(i)}\middle|\begin{bmatrix}\mu^*_{1\Phi}(\boldsymbol{v}^{(i)})\\ \vdots \\ \mu^*_{p\Phi}(\boldsymbol{v}^{(i)})\end{bmatrix}, \operatorname{diag}\begin{bmatrix}\sigma^*_{1\Phi}(\boldsymbol{v}^{(i)})\\ \vdots \\ \sigma^*_{p\Phi}(\boldsymbol{v}^{(i)})\end{bmatrix}\right)}_{\text{approximating } \mathcal{B}\left(\boldsymbol{v}^{(i)}|, p_\theta(\boldsymbol{y}^{(i)})\right)} \mathcal{N}\left(\boldsymbol{y}^{(1:n_f)}\middle|\mathbf{0}, K\right) \\ \text{with } K &= \begin{bmatrix} K_{11}(T,T) & \dots & K_{1p}(T,T) \\ \vdots & & \vdots \\ K_{1p}(T,T)^\top & \dots & K_{pp}(T,T) \end{bmatrix} \qquad (10) \end{aligned}$$

that is based on the model (5) but with a Gaussian approximation $q^*_\Phi(\boldsymbol{y}^{(i)}|\boldsymbol{v}^{(i)})$ of the Bernoulli term $\mathcal{B}$ in (5) that represents the pixel model. Since the Gaussian distribution is conjugate to itself, the approximation allows us to obtain the exact posterior distribution. The $\boldsymbol{y}^{(1:n_f)}$ are latent function values, and $\{(t_i, \mu^*_{j\Phi}(\boldsymbol{v}^{(i)}))\}_{i=1}^{n_f}$ are a set of pseudo-inputs $\mu^*_{j\Phi}(\boldsymbol{v}^{(i)}) \in \mathbb{R}$ each with

noise $\sigma^*_{j\Phi}(\boldsymbol{v}^{(i)}) \in \mathbb{R}_+$ for $j \in \{1, \dots, p\}$ provided by the encoder network. Conditioning the GP prior on these points leads to an analytically tractable posterior that approximates the true posterior $p(\boldsymbol{y}^{(1:n_f)}|\boldsymbol{v}^{(1:n_f)})$. The function $L(\boldsymbol{v}^{(1:n_f)})$ is the standard marginal likelihood of the GP, see Rasmussen and Williams (2006), given by

$$\log L(\boldsymbol{v}^{(1:n_f)}) = -\frac{1}{2}\left([\boldsymbol{\mu}^*]^\top (K + \Sigma^*)^{-1}\boldsymbol{\mu}^* - \log|K + \Sigma^*| - n_f \log 2\pi\right) \tag{11}$$
$$\text{with } \boldsymbol{\mu}^* = \begin{bmatrix} \mu^*_1(\boldsymbol{v}^{(1:n_f)}) \\ \vdots \\ \mu^*_p(\boldsymbol{v}^{(1:n_f)}) \end{bmatrix}, \Sigma^* = \operatorname{diag}\begin{bmatrix} \sigma^*_{1\Phi}(\boldsymbol{v}^{(1:n_f)}) \\ \vdots \\ \sigma^*_{p\Phi}(\boldsymbol{v}^{(1:n_f)}) \end{bmatrix},$$

which is typically used to optimized the kernel's hyperparameters. In the next section, we present the training of the PEGP-VAE.

3.3. Training

Learning and inference for the PEGP-VAE are concerned with determining the parameters of the encoder $\boldsymbol{\Phi}$, the parameters of the decoder $\boldsymbol{\theta}$, and the unknown parameters in the latent dynamics $\boldsymbol{\varphi}$. For this purpose, we are maximizing the evidence lower bound (ELBO) given by

$$\mathcal{L}_{ELBO}(\boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\varphi}, \boldsymbol{v}^{(1:n_f)}) = \mathrm{E}\left[\sum_{i=1}^{n_f} \log \mathcal{B}\big(\boldsymbol{v}^{(i)}|p_\theta(\boldsymbol{y}^{(i)}) - \log q^*_\Phi(\boldsymbol{y}^{(i)}|\boldsymbol{v}^{(i)})\right] + \log L(\boldsymbol{v}^{(1:n_f)}).$$

The first term is the reconstruction term, evaluated with the reparameterization trick (Kingma and Welling, 2013), which must be evaluated by Monte-Carlo sampling. The middle and the right term compose the analytically tractable Kullback-Leibler divergence between the GP prior and the inference model. Alternatively, the first two terms together may be viewed as the error between the true posterior and approximate posterior, since the Bernoulli likelihoods are approximated by a Gaussian distribution. Finally, the last term is the log marginal likelihood (11) of the GP. For more information on the ELBO function see Pearce (2020).

4. Simulation

Setting: To highlight the benefits of the proposed PEGP-VAE, we consider observing a micro-particle in a 2-dimensional space. The particle is excited by an unknown, time-dependent electromagnetic field. We assume that we know the resonance frequency and damping factor of the particle such that we assume an harmonic oscillator as prior knowledge on the latent dynamics given by

$$\dot{\boldsymbol{x}}(t) = \underbrace{\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -c_1 & 0 & -d_1 & 0 \\ 0 & -c_2 & 0 & -d_2 \end{bmatrix}}_{A} \boldsymbol{x}(t) + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}}_{B} \boldsymbol{u}(t), \quad \boldsymbol{y}(t) = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{C} \boldsymbol{x}(t) \tag{12}$$

with the electromagnetic field input $\boldsymbol{u} \in \mathbb{R}^2$. Here, the latent state y_1 describes the horizontal position, while y_2 is the vertical position, i.e., the dimension of the latent space is $q = 2$. The

constants c_1, c_2, d_1, d_2 are selected such that the particle has a resonance frequency of $47.7\,\text{kHz}$ / $63.6\,\text{kHz}$ and a damping factor of 0.02 / 0.01 for the horizontal and vertical direction, respectively. 100 video sequences of particles with a resolution of 40×40 pixels are artificially generated as training data using samples from a GP prior with squared exponential kernel for the input u_1, u_2 of (12). Each video sequence has a duration of $30\,\mu\text{s}$ with one frame per μs. In Figure 2, four examples of generated particle movements are shown.

Figure 2: Four examples of particle movement sequences out of the training set as heatmaps (increasing brightness from start to end).

Configuration: The GP prior on the latent state is equipped with the physics-enhanced kernel (8) with a squared exponential kernel as prior for the inputs u_1, u_2 and the Green's function of (12) given by $G(t, t') = C\boldsymbol{e}^{A(t-t')}B \in \mathbb{R}^{2\times 2}$. In Figure 3, the correlation between two points in time for the squared exponential kernel (left) and the physics-enhanced kernel (right) is shown. The periodicity and damping of the oscillator manifest themselves as a repetitive, decreasing correlation over time.

For the input encoder, we use a fully connected network that takes a frame $\boldsymbol{v}^{(i)} \in \{0,1\}^{40\cdot 40}$ of the video sequence as input. The input layer is followed by a fully connected hidden layer of 500 nodes with a $\tanh$-activation function, and the output layer consisting of four nodes returning the pseudo-inputs $\mu^*_{1\Phi}(\boldsymbol{v}^{(i)}), \mu^*_{2\Phi}(\boldsymbol{v}^{(i)})$ and noise $\log\sigma^*_{1\Phi}(\boldsymbol{v}^{(i)}), \log\sigma^*_{2\Phi}(\boldsymbol{v}^{(i)})$. Thus, the network is parametrized by two weight matrices W^1_Φ, W^2_Φ and two bias vectors B^1_Φ, B^2_Φ such that $\Phi = \{W^1_\Phi, B^1_\Phi, W^2_\Phi, B^2_\Phi\}$. Analogously, the decoder consists of an input layer with $p = 2$ inputs, a fully connected hidden layer of 500 nodes with the $\tanh$-activation and $40 \cdot 40 = 1600$ nodes with the sigmoid-activation function to achieve a Bernoulli probability between zero and one for each pixel.

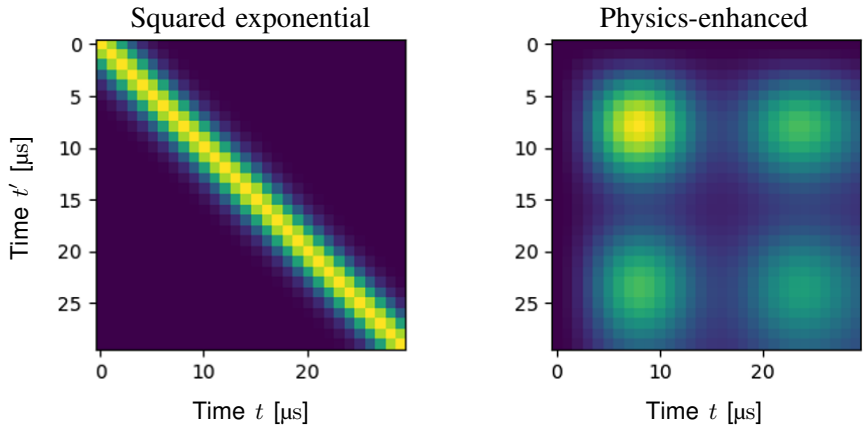

Figure 3: Correlation of two points in time for the squared exponential kernel and the physics-enhanced kernel that represents a damped oscillator.

The decoder network is parameterized by $\boldsymbol{\theta} = \{W_\theta^1, B_\theta^1, W_\theta^2, B_\theta^2\}$. The training (maximization of the ELBO) is implemented in Python using PyTorch and the Adam optimizer with a learning rate of $1e-3$. Each method is trained for 30000 iterations.

Results: Figure 4 show the reconstructed video sequences for two samples of the test set. On the left side, the original video sequences are visualized. The videos are used as input for the trained encoder network, and the resulting GP posterior for the latent state over time is shown in the second column. The crosses mark the unknown ground truth. The GP with a physics-enhanced kernel is able to reconstruct the unknown trajectory of the latent state. Furthermore, all samples of the GP are respecting the latent dynamics (12). On the right side of Figure 4, the reconstructed videos using the latent state trajectory as input for the trained decoder network are depicted. We assume that the quality of the decoder can be even further improved by more hidden nodes in the neural network and/or more training data.

In Figure 5, we compare the reconstruction quality of the latent state for a GP with squared exponential kernel (left) against the physics-enhanced kernel (right) for the two samples shown in Fig. 4 (top and bottom, respectively). In this case, the input video sequence has a duration of $30\,\mu$s (black line), and we aim to predict a video sequence for $50\,\mu$s. Both VAE are trained for the same number of iterations. The unknown ground truth is marked by crosses. The physics-enhanced kernel clearly outperforms the squared exponential kernel in terms of reconstruction accuracy and generalization quality. Although the uncertainty (shaded area) for both approaches increases after $30\,\mu$s,

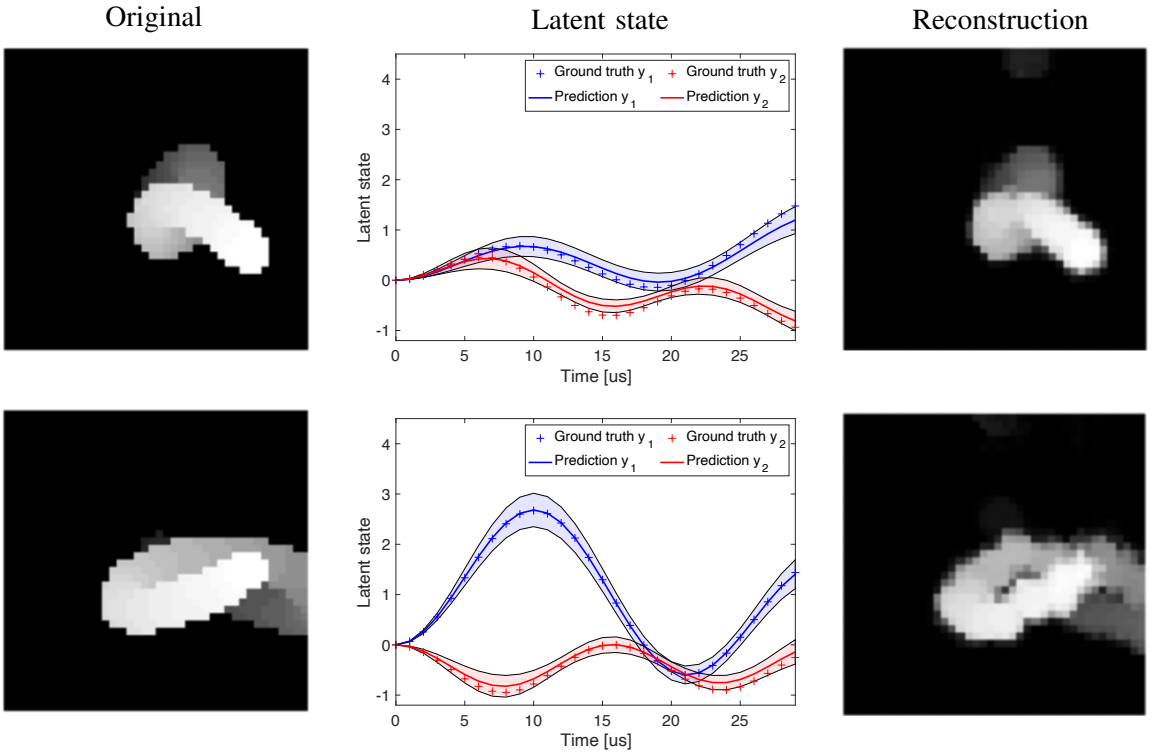

Figure 4: Left: Original videos. Middle: Horizontal y_1 (blue) and vertical position y_2 (red) of the particle. The mean prediction of the latent state (solid line) and 2σ-uncertainty (shaded area) of the GP with a physics-enhanced kernel. The crosses are the unknown ground truth. Right: Reconstructed videos.

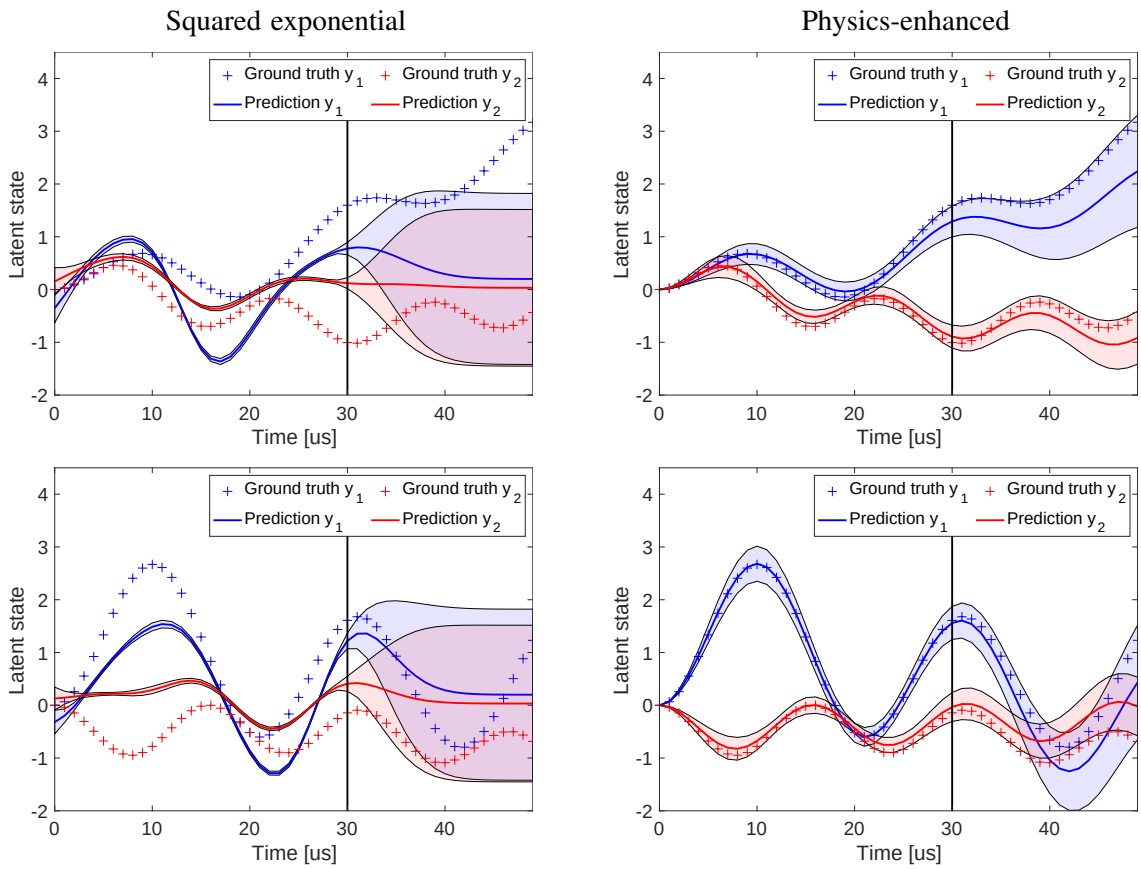

Figure 5: Comparison of GP based VAE with squared exponential kernel against physics-enhanced kernel for two samples (top/bottom row) over a horizon of 50 µs. The horizontal line at 30 µs marks the end of the training sequences. The physics-enhanced kernel is superior in terms of accuracy and generalization of the latent state.

the PEGP-VAE benefits from the encoded prior knowledge, whereas the squared exponential kernel performs poorly on the previously unseen time interval. Due to the reduced uncertainty using the physics-enhanced kernel, we also observe a significant improvement in the reconstruction of the trajectory.

Conclusion

We propose a physics-enhanced Gaussian process variational autoencoder (PEGP-VAE) for learning physically correct latent dynamics from pixels. For this purpose, we place a GP prior on the latent time series, where the GP is based on a physics-enhanced kernel. This kernel is derived using latent force models and the Green's function of the physical model expressed by linear dynamics. The proposed approach improves the reconstruction quality of the latent state as the space of potential latent dynamics is reduced and respects physical prior knowledge. For future work, we plan to use convolutional NN for the encoder/decoder due to the spatio-temporal nature of the data.

References

Mauricio A. Alvarez, David Luengo, and Neil D. Lawrence. Linear latent force models using Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11): 2693–2705, 2013.

Alexander Amini, Wilko Schwarting, Guy Rosman, Brandon Araki, Sertac Karaman, and Daniela Rus. Variational autoencoder for end-to-end control of autonomous driving with novelty detection and training de-biasing. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 568–575, 2018.

Matthew Ashman, Jonathan So, Will Tebbutt, Vincent Fortuin, Michael Pearce, and Richard E Turner. Sparse Gaussian process variational autoencoders. *arXiv preprint arXiv:2010.10177*, 2020.

David Barber, A. Taylan Cemgil, and Silvia Chiappa. *Bayesian time series models*. Cambridge University Press, 2011.

Christopher M. Bishop and Nasser M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Alex Campbell and Pietro Liò. tvGP-VAE: Tensor-variate Gaussian process prior variational autoencoder. *arXiv preprint arXiv:2006.04788*, 2020.

Francesco Paolo Casale, Adrian V Dalca, Luca Saglietti, Jennifer Listgarten, and Nicolo Fusi. Gaussian process prior variational autoencoders. *arXiv preprint arXiv:1810.11738*, 2018.

Richard Courant and David Hilbert. *Methods of mathematical physics: partial differential equations*. John Wiley & Sons, 2008.

Emmanuel De Bézenac, Arthur Pajot, and Patrick Gallinari. Deep learning for physical processes: Incorporating prior scientific knowledge. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124009, 2019.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

N Benjamin Erichson, Michael Muehlebach, and Michael W Mahoney. Physics-informed autoencoders for Lyapunov-stable fluid flow prediction. *arXiv preprint arXiv:1905.10866*, 2019.

Marco Farina, Yuichiro Nakai, and David Shih. Searching for new physics with deep autoencoders. *Physical Review D*, 101(7):075021, 2020.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. Gp-vae: Deep probabilistic time series imputation. In *International Conference on Artificial Intelligence and Statistics*, pages 1651–1661. PMLR, 2020.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. *arXiv preprint arXiv:1710.05741*, 2017.

Andreas Geist and Sebastian Trimpe. Learning constrained dynamics with Gauss' principle adhering Gaussian processes. In *Learning for Dynamics and Control*, pages 225–234. PMLR, 2020.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pages 1462–1471. PMLR, 2015.

Adi Hanuka, Xiaobiao Huang, Jane Shtalenkova, Dylan Kennedy, Auralee Edelen, VR Lalchand, Daniel Ratner, and Joseph Duris. Physics-informed Gaussian process for online optimization of particle accelerators. *arXiv preprint arXiv:2009.03566*, 2020.

Zhong-Sheng Hou and Zhuo Wang. From model-based control to data-driven control: Survey, classification and perspective. *Information Sciences*, 235:3–35, 2013.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Wu Lin, Nicolas Hubacher, and Mohammad Emtiyaz Khan. Variational message passing with structured inference networks. *arXiv preprint arXiv:1803.05589*, 2018.

Ilia A. Luchnikov, Alexander Ryzhov, Pieter-Jan Stas, Sergey N Filippov, and Henni Ouerdane. Variational autoencoder reconstruction of complex many-body physics. *Entropy*, 21(11):1091, 2019.

Masatoshi Nagano, Tomoaki Nakamura, Takayuki Nagai, Daichi Mochihashi, and Ichiro Kobayashi. Spatio-temporal categorization for first-person-view videos using a convolutional variational autoencoder and Gaussian processes. *Frontiers in Robotics and AI*, 9, 2022.

Michael Pearce. The Gaussian process prior VAE for interpretable latent dynamics from pixels. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–12. PMLR, 2020.

Michael Pearce, Silvia Chiappa, and Ulrich Paquet. Comparing interpretable inference models for videos of physical motion. In *1st Symposium on Advances in Approximate Bayesian Inference*, 2018.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.

Lucas Rath, Andreas René Geist, and Sebastian Trimpe. Using physics knowledge for learning rigid-body forward dynamics with Gaussian process force priors. In *5th Annual Conference on Robot Learning*, 2021.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286. PMLR, 2014.

Vaibhav Saxena, Jimmy Ba, and Danijar Hafner. Clockwork variational autoencoders. *arXiv preprint arXiv:2102.09532*, 2021.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, 28: 3483–3491, 2015.

Mark Van der Wilk, Carl Edward Rasmussen, and James Hensman. Convolutional Gaussian processes. *arXiv preprint arXiv:1709.01894*, 2017.

Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, pages 835–851. Springer, 2016.

Zheng Wang, Wei Xing, Robert Kirby, and Shandian Zhe. Physics regularized Gaussian processes. *arXiv preprint arXiv:2006.04976*, 2020.

Faraz Waseem, Rafael Perez Martinez, and Chris Wu. Visual anomaly detection in video by variational autoencoder. *arXiv preprint arXiv:2203.03872*, 2022.