

---

# Learning deep dynamical models from image pixels

---

Niklas Wahlström<sup>1</sup>

Thomas B. Schön<sup>2</sup>

Marc Peter Deisenroth<sup>3</sup>

<sup>1</sup>Department of Electrical Engineering, Linköping University, Sweden

<sup>2</sup>Department of Information Technology, Uppsala University, Sweden

<sup>3</sup>Department of Computing, Imperial College London, United Kingdom

## Abstract

Modeling dynamical systems is important in many disciplines, e.g., control, robotics, or neurotechnology. Commonly the state of these systems is not directly observed, but only available through noisy and potentially high-dimensional observations. In these cases, system identification, i.e., finding the measurement mapping and the transition mapping (system dynamics) in latent space can be challenging. For linear system dynamics and measurement mappings efficient solutions for system identification are available. However, in practical applications, the linearity assumptions does not hold, requiring non-linear system identification techniques. If additionally the observations are high-dimensional (e.g., images), non-linear system identification is inherently hard. To address the problem of non-linear system identification from high-dimensional observations, we combine recent advances in deep learning and system identification. In particular, we jointly learn a low-dimensional embedding of the observation by means of deep auto-encoders and a predictive transition model in this low-dimensional space. We demonstrate that our model enables learning good predictive models of dynamical systems from pixel information only.

## 1 Introduction

High-dimensional time series include video streams, Electroencephalography (EEG) and sensor network

data. Dynamical models describing such data are desired for forecasting (prediction) and controller design, both of which play an important role, e.g., in autonomous systems, machine translation, robotics and surveillance applications. A key challenge is system identification, i.e., finding a mathematical model of the dynamical system based on the information provided by measurements from the underlying system. In the context of state-space models this includes finding two functional relationships between (a) the states at different time steps (prediction/transition model) and (b) states and corresponding measurements (observation/measurement model). In the linear case, this problem is very well studied, and many standard techniques exist, e.g., subspace methods [18], expectation maximization [16, 4] and prediction-error methods [9]. However, in realistic and practical scenarios we require non-linear system identification techniques.

Learning non-linear dynamical models is an inherently difficult problem, and it has been one of the most active areas in system identification for the last decades [10, 17]. In recent years, sequential Monte Carlo (SMC) methods have received attention for identifying non-linear state-space models [15], see also the recent surveys [6]. While methods based on SMC are powerful, they are also computationally expensive. Learning non-linear dynamical models from very high-dimensional sensor data is even more challenging. First, finding (non-linear) functional relationships in very high dimensions is hard (un-identifiability, local optima, overfitting, etc.); second, the amount of data required to find a good function approximator is enormous. Fortunately, high-dimensional data often possesses an intrinsic lower dimensionality. We will exploit this property for system identification by finding a low-dimensional representation of high-dimensional data and learning predictive models in this low-dimensional space. For this purpose, we need an automated procedure for finding compact low-dimensional representations/features.

The state of the art in learning parsimonious representations of high-dimensional data is currently defined by deep learning architectures, such as deep neural networks [5], stacked/deep auto-encoders [19] and convolutional neural networks [8], all of which have been successfully applied to image, text, speech and audio data in commercial products, e.g., by Google, Amazon and Facebook. Typically, these feature learning methods are applied to static data sets, e.g., for image classification. The auto-encoder gives explicit expressions of two generative mappings: 1) an encoder  $g^{-1}$  mapping the high-dimensional data to the features, and 2) a decoder  $g$  mapping the features to high-dimensional reconstructions. In the machine learning literature, there exists a vast number of other well studied nonlinear dimensionality reduction methods such as the Gaussian process latent variable model (GP-LVM) [7], kernel PCA [14], Laplacian Eigenmaps [1] and Locally Linear Embedding [12]. However, all of them provide at most one of the two mappings  $g$  and  $g^{-1}$ .

In this paper, we combine feature learning and dynamical systems modeling to obtain good predictive models for high-dimensional time series. In particular, we use deep auto-encoder neural networks for automatically finding a compact low-dimensional representation of an image. In this low-dimensional feature space, we use a neural network for modeling the nonlinear system dynamics. The embedding and the predictive model in feature space are learned *jointly*. An simplified illustration of our approach is shown in Fig. 1. An encoder  $g^{-1}$  maps an image  $y_{t-1}$  at time step  $t-1$  to a low-dimensional feature  $z_{t-1}$ . In this feature space, a prediction model  $l$  maps the feature forward in time to  $z_t$ . Subsequently, the decoder  $g$  can be used to generate a predicted image  $y_t$  at the next time step. This framework needs access to both the encoder  $g^{-1}$  and the decoder  $g$ , which motivates our use of the auto-encoder as dimensionality reduction technique.

Consequently, the contributions of this paper are (a) a model for learning a low-dimensional dynamical representation of high-dimensional data, which can be used for long-term predictions; (b) experimental evidence demonstrating that jointly learning the parameters of the latent embedding and the predictive model in latent space can increase the performance compared to a separate training.

## 2 Model

We consider a dynamical system where control inputs are denoted by  $u$  and observations are denoted by  $y$ . In the context of this paper, the observations are pixel information from images. We assume that

a low-dimensional latent variable  $z$  exists that compactly represents the relevant properties of  $y$ . Since we consider dynamical systems, a low-dimensional representation  $z$  of a (static) image  $y$  is insufficient to capture important dynamic information, such as velocities. Therefore, we introduce an additional latent variable  $x$ , the *state*. In our case, the state  $x_t$  contains features from multiple time steps (e.g.,  $t-1$  and  $t$ ) to capture velocity (or higher-order) information. Therefore, our transition model does not map features at time  $t-1$  to time  $t$  (as illustrated in Fig. 1), but the transition function  $f$  maps states  $x_{t-1}$  (and controls  $u_{t-1}$ ) to states  $x_t$  at time  $t$ . The full dynamical system is given as the state-space model

$$x_{t+1} = f(x_t, u_t; \theta) + w_t(\theta), \quad (1a)$$

$$z_t = h(x_t; \theta) + v_t(\theta), \quad (1b)$$

$$y_t = g(z_t; \theta) + e_t(\theta), \quad (1c)$$

where each measurement  $y_t$  can be described by a low-dimensional feature representation  $z_t$  (1c). These features are in turn modeled with a low-dimensional state-space model in (1a) and (1b), where the state  $x_t$  contains the full information about the state of the system at time instant  $t$ , see also Fig. 2a. Here  $w_t(\theta)$ ,  $v_t(\theta)$  and  $e_t(\theta)$  are sequences of independent random variables and  $\theta$  are the model parameters.

### 2.1 Approximate predictor model

To identify parameters in dynamical systems, the prediction-error method [9] has been applied extensively within the system identification community during the last five decades. It is based on minimizing the error between the sequence of measurements  $y_t$  and the predictions  $\hat{y}_{t|t-1}(\theta)$ , usually the one-step ahead prediction. To achieve this, we need a predictor model that relates the prediction  $\hat{y}_{t|t-1}(\theta)$  to all previous measurements, control inputs and the system parameters  $\theta$ .

In general, it is difficult to derive a predictor model based on the nonlinear state-space model (1), and a closed form expression for the prediction is only available in a few special cases [9]. However, by approximating the optimal solution, a predictor model for the features  $z_t$  can be stated in the form

$$\hat{z}_{t|t-1}(\theta_M) = l(Z_{t-1}; \theta_M), \quad (2)$$

where  $Z_{t-1} = (z_1, u_1, \dots, z_{t-1}, u_{t-1})$  includes all past features and control inputs,  $l$  is a nonlinear function and  $\theta_M$  is the corresponding model parameters. Note that the predictor model no longer has an explicit notion of the state  $x_t$ . The model introduced in (2) is indeed very flexible, and in this work we have restricted this flexibility somewhat by working with a

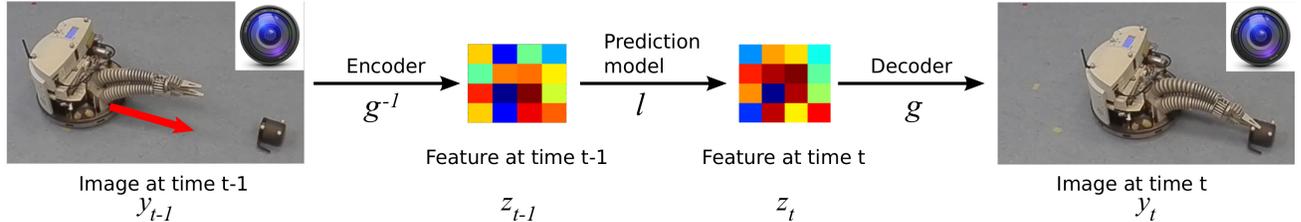


Figure 1: Combination of deep learning architectures for feature learning and predictor models in feature space. A camera observes a robot approaching an object. A good low-dimensional feature representation of an image is important for learning a predictive model if the camera is the only sensor available.

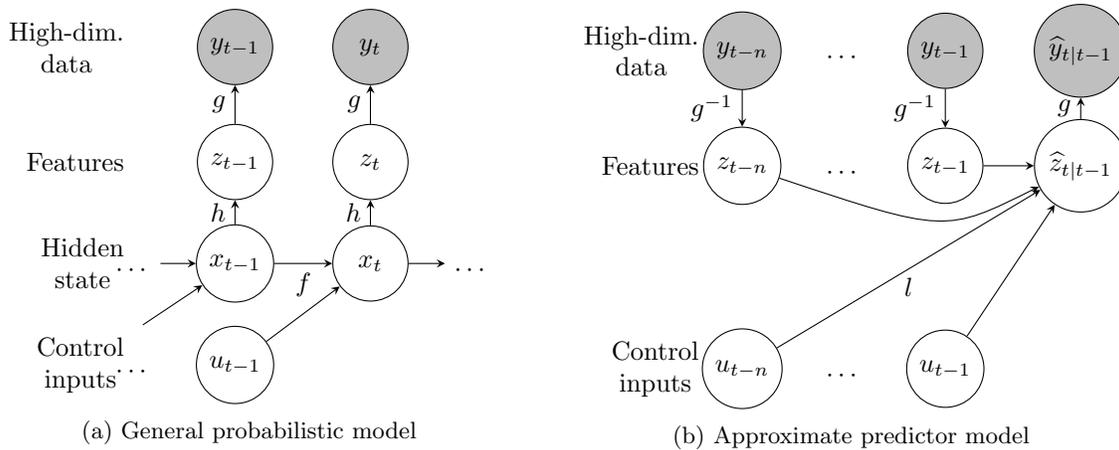


Figure 2: (a) The general graphical model of the high-dimensional data  $y_t$ . Each data point  $y_t$  has a low-dimensional representation  $z_t$ , which is modeled using a state-space model with hidden state  $x_t$  and control input  $u_t$ . (b) The approximate predictor model, where the predicted feature  $\hat{z}_{t|t-1}$  is a function of the  $n$  past features  $z_{t-n}$  to  $z_{t-1}$  and  $n$  past control inputs  $u_{t-n}$  to  $u_{t-1}$ . Each of the features  $z_{t-n}$  to  $z_{t-1}$  is computed from high-dimensional data  $y_{t-n}$  to  $y_{t-1}$  via the encoder  $g^{-1}$ . The predicted feature  $\hat{z}_{t|t-1}$  is mapped to predicted high-dimensional data via the decoder  $g$ .

nonlinear autoregressive exogenous model (NARX) [9], which relates the predicted current value  $\hat{z}_{t|t-1}(\theta_M)$  of the time series to the past  $n$  values of the time series  $z_{t-1}, z_{t-2}, \dots, z_{t-n}$ , as well as the past  $n$  values of the control inputs  $u_{t-1}, u_{t-2}, \dots, u_{t-n}$ . The resulting NARX predictor model is given by

$$\hat{z}_{t|t-1}(\theta_M) = l(z_{t-1}, u_{t-1}, \dots, z_{t-n}, u_{t-n}; \theta_M), \quad (3)$$

where  $l$  is a nonlinear function, in our case a neural network. The model parameters in the nonlinear function are normally estimated by minimizing the sum of the prediction errors  $z_t - \hat{z}_{t|t-1}(\theta_M)$ . However, since we are interested in a good predictive performance for the high-dimensional data  $y$  rather than for the features  $z$ , we transform the predictions back to the high-dimensional space and obtain a prediction  $\hat{y}_{t|t-1} = g(\hat{z}_{t|t-1}; \theta_D)$ , which we use in our error measure.

An additional complication is that we do not have access to the features  $z_t$ . Therefore, before training, the past values of the time series have to be replaced with

their feature representation  $y = g^{-1}(y; \theta_E)$ , which we compute from the pixel information  $y$ . Here,  $g^{-1}$  is an approximate inverse of  $g$ , which will be described in more detail the next section. This gives the final predictor model

$$\hat{y}_{t|t-1}(\theta_E, \theta_D, \theta_M) = g(\hat{z}_{t|t-1}(\theta_E, \theta_M); \theta_D), \quad (4a)$$

$$\hat{z}_{t|t-1}(\theta_E, \theta_M) = l(z_{t-1}(\theta_E), u_{t-1}, \dots, z_{t-n}(\theta_E), u_{t-n}; \theta_M), \quad (4b)$$

$$z_t(\theta_E) = g^{-1}(y_t; \theta_E), \quad (4c)$$

which is also illustrated in Fig. 2b. The corresponding prediction error will be

$$\varepsilon_t^P(\theta_E, \theta_D, \theta_M) = y_t - \hat{y}_{t|t-1}(\theta_E, \theta_D, \theta_M). \quad (5)$$

## 2.2 Auto-encoder

We use a deep auto-encoder neural network to parameterize the feature mapping and its inverse. It consists of a deep encoder network  $g^{-1}$  and a deep decoder network  $g$ . Each layer  $k$  of the encoder neural network

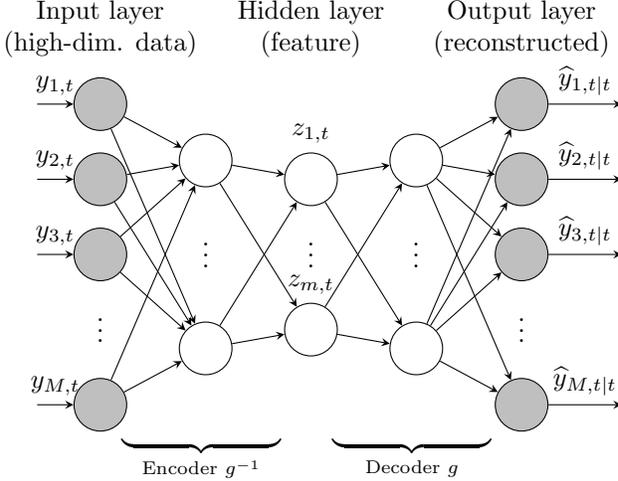


Figure 3: An auto-encoder consisting of an encoder  $g^{-1}$  and a decoder  $g$ . The original image  $y_t = [y_{1,t}, \dots, y_{M,t}]^T$  is mapped to its low-dimensional representation  $z_t = [z_{1,t}, \dots, z_{m,t}]^T = g^{-1}(y_t)$  with the encoder, and then back to a high-dimensional representation  $\hat{y}_{t|t-1} = g(\hat{z}_{t|t-1})$  by the decoder  $g$ , where  $M \gg m$ .

$g^{-1}$  computes  $y_t^{(k+1)} = \sigma(A_k y_t^{(k)} + b_k)$ , where  $\sigma$  is a squashing function and  $A_k$  and  $b_k$  are free parameters. The control input to the first layer is the image, i.e.,  $y_t^{(1)} = y_t$ . The last layer is the low-dimensional feature representation of the image  $z_t(\theta_E) = g^{-1}(y_t; \theta_E)$ , where  $\theta_E = [\dots, A_k, b_k, \dots]$  are the parameters of all neural network layers. The decoder  $g$  consists of the same number of layers in reverse order, see Fig. 3. It can be considered an approximate inverse of the encoder, such that  $\hat{y}_{t|t}(\theta_E, \theta_D) \approx y_t$ , where

$$\hat{y}_{t|t}(\theta_E, \theta_D) = g(g^{-1}(y_t; \theta_E); \theta_D) \quad (6)$$

is the reconstructed version of  $y_t$ . In the classical setting, the encoder and the decoder are trained simultaneously to minimize the reconstruction error

$$\varepsilon_t^R(\theta_E, \theta_D) = y_t - \hat{y}_{t|t}(\theta_E, \theta_D), \quad (7)$$

where the parameters of  $g$  and  $g^{-1}$  optionally can be coupled to constrain the solution to some degree [19].

We realize that the autoencoder suits our problem at hand very well, since it provides an explicit expression of both the mapping from features to data  $g$  as well as its approximate inverse  $g^{-1}$ , which is convenient to form the predictions in (4a). Many other nonlinear dimensionality reduction methods such as the Gaussian process latent variable model (GP-LVM) [7], kernel PCA [14], Laplacian Eigenmaps [1] and Locally Linear Embedding [12] do not provide an explicit expression of both mappings  $g$  and  $g^{-1}$ . This is the main

motivation way we use the (deep) auto-encoder model for dimensionality reduction.

### 3 Training

To summarize, our model contains the following free parameters: the parameters for the encoder  $\theta_E$ , the parameters for the decoder  $\theta_D$  and the parameters for the predictor model  $\theta_M$ . To train the model, we employ two cost functions, the sum of the prediction errors (5),

$$V_P(\theta_E, \theta_D, \theta_M) = \log \left( \frac{1}{2NM} \sum_{t=1}^N \|\varepsilon_t^P(\theta_E, \theta_D, \theta_M)\|^2 \right), \quad (8a)$$

and the sum of the reconstruction errors (7),

$$V_R(\theta_E, \theta_D) = \log \left( \frac{1}{2NM} \sum_{t=1}^N \|\varepsilon_t^R(\theta_E, \theta_D)\|^2 \right). \quad (8b)$$

Generally, there are two ways of finding the model parameters: (1) separate training and (2) joint training, both of which are explained below.

#### 3.1 Separate training

Normally when features are used for inference of dynamical models, they are first extracted from the data in a pre-processing step and as a second step the predictor model is estimated based on these features. In our setting, this would correspond to sequentially training the model with using two cost functions (8a)–(8b) where we first learn a compact feature representation by minimizing the reconstruction error

$$(\hat{\theta}_E, \hat{\theta}_D) \in \arg \min_{\theta_E, \theta_D} V_R(\theta_E, \theta_D), \quad (9a)$$

and subsequently minimize the prediction error

$$\hat{\theta}_M = \arg \min_{\theta_M} V_P(\hat{\theta}_E, \hat{\theta}_D, \theta_M) \quad (9b)$$

with fixed auto-encoder parameters  $\hat{\theta}_E, \hat{\theta}_D$ . The gradients of these cost functions with respect to the model parameters can be computed efficiently by back-propagation. The cost functions are then minimized by the BFGS algorithm [11].

#### 3.2 Joint training

An alternative to separate training is to minimize the reconstruction error and the prediction error simultaneously by considering the optimization problem

$$(\hat{\theta}_E, \hat{\theta}_D, \hat{\theta}_M) = \arg \min_{\theta_E, \theta_D, \theta_M} (V_R(\theta_E, \theta_D) + V_P(\theta_E, \theta_D, \theta_M)), \quad (10)$$

where we jointly optimize the free parameters in both the auto-encoder  $\theta_E$ ,  $\theta_D$  and the predictor model  $\theta_M$ . Again, back-propagation is used for computing the gradients of this cost function.

### 3.3 Initialization

The auto-encoder has strong similarities with principal component analysis (PCA). More precisely, if we use a linear activation function and only consider a single layer, the auto-encoder and PCA are identical [2]. We exploited this relationship to initialize the parameters of the auto-encoder. The auto-encoder network has been unfolded, each pair of layers in the encoder and the decoder have been combined, and the corresponding PCA solution has been computed for each of these pairs. By starting with the high-dimensional data at the top layer and using the principal components from that pair of layers as input to the next pair of layers, we recursively compute a good initialization for all parameters in the auto-encoder network. Similar pre-training routines are found in [5] where a restricted Boltzmann machine is used instead of PCA.

## 4 Results

We report results on identification of (1) the nonlinear dynamics of a pendulum (1-link robot arm) moving in a horizontal plane and the torque as control input, (2) an object moving in the 2D-plane, where the 2D velocity serves as control input. In both examples, we learn the dynamics solely based on pixel information. Each pixel  $y_t^{(i)}$  is a component of the measurement  $y_t = [y_t^{(1)}, \dots, y_t^{(M)}]^\top$  and assumes a continuous gray-value in  $[0, 1]$ .

### 4.1 Experiment 1: Pendulum in the plane

We simulated 400 frames of a pendulum moving in a plane with  $51 \times 51 = 2601$  pixels in each frame and the torque of the pendulum as control input. To speed up training, the image input has been reduced to  $\dim(y_t) = 50$  prior to model learning (system identification) using PCA. With these 50 dimensional inputs, four layers have been used for the encoder  $g^{-1}$  as well as the decoder  $g$  with dimension 50-25-12-6-2. Hence, the features have dimension  $\dim(x_t) = 2$ . The order of the dynamics was chosen as  $n = 2$  to capture velocity information. For the predictor model  $l$  we used a two-layer neural network with a 6-4-2 architecture.

We evaluate the performance in terms of long term predictions. These predictions are constructed by concatenating multiple 1-step ahead predictions. More precisely, the  $p$ -step ahead prediction  $\hat{y}_{t+p|t} = g(\hat{z}_{t+p|t})$

Table 1: Exp. 1: Prediction error  $V_P$  and reconstruction error  $V_R$  for separate and joint training.

Training	$V_P$	$V_R$
Joint training (10)	-6.91	-6.92
Separate training (9)	-5.12	-6.99

is computed iteratively as

$$\hat{z}_{t+1|t} = l(\hat{z}_{t|t}, u_t, \dots), \quad (11a)$$

...

$$\hat{z}_{t+p|t} = l(\hat{z}_{t+p-1|t}, u_{t+p-1|t}, \dots), \quad (11b)$$

where  $\hat{z}_{t|t} = g^{-1}(y_t)$  is the features of the data point at time instance  $t$ . We assumed that the applied future torques were known.

The predictive performance on two exemplary image sequences of the validation data of our system identification models is illustrated in Fig. 4, where control inputs (torques) are assumed known. The top rows show the ground truth images, the center rows show the predictions based on a model using joint training (10), the bottom rows show the corresponding predictions of a model where the auto-encoder and the predictive model were trained sequentially according to (9). For the model based on jointly training all parameters, we obtain good predictive performance for both one-step ahead prediction and multiple-step ahead prediction. In contrast, the predictive performance of learning the features and the dynamics separately is worse than the predictive performance of the model trained by jointly optimizing all parameters. Although the auto-encoder does a perfect job (left-most frame, 0-step ahead prediction), already the (reconstructed) one-step ahead prediction is not similar to the ground-truth image. This can also be seen in Table 1 where the reconstruction error is equally good for both models, but for the prediction error we manage to get a better value using joint training than using separate training. Let us have a closer look at the model based on separate training: Since the auto-encoder performs well, the learned transition model is the reason for bad predictive performance. We believe that the auto-encoder found a good feature representation for reconstruction, but this representation was not ideal for learning a transition model.

Fig. 5 displays the “decoded” images corresponding to the latent representations using joint and separate training, respectively. In the joint training the feature values line up in a circular shape enabling a low-dimensional dynamical description, whereas separate training finds feature values, which are not even placed sequentially in the low-dimensional representation. Separate training extracts the low-dimensional representations without context, i.e., the knowledge

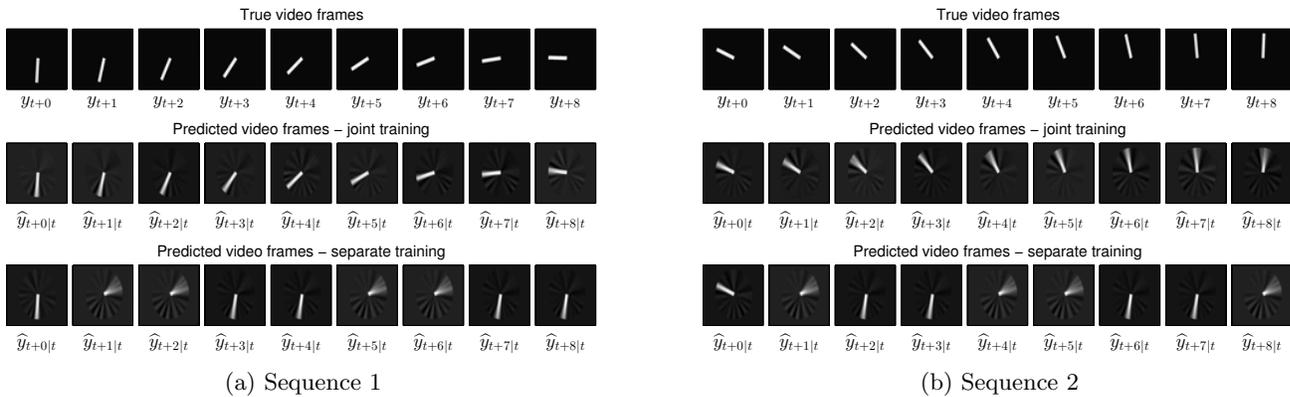


Figure 4: Exp. 1: Two typical image sequences and corresponding prediction results (validation data), computed according to (11). The top rows show nine consecutive ground truth image frames from time instant  $t$  to  $t + 8$ . The second and the third rows display the corresponding long-term ahead predictions based on measured images up to time  $t$  for both joint (center) and separate training (bottom) of the model parameters.

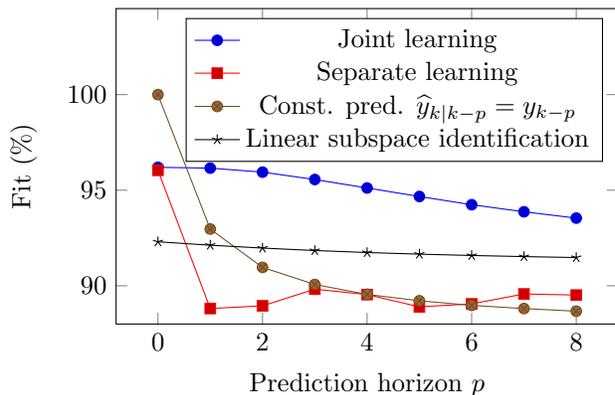


Figure 6: Exp. 1: Fitting quality (12) for joint and separate learning of features and dynamics for different prediction horizons  $p$ . The fit is compared with the naive prediction  $\hat{y}_{t|t-p} = y_{t-p}$ , where the most recent image is used for the prediction  $p$  steps ahead and a linear subspace-ID method.

that these features constitute a time-series. On the other hand, joint training enables the extraction of features that can also model the dynamical behavior in a compact manner.

In this particular data set, the data points clearly reside on one-dimensional manifold, encoded by the pendulum angle. However, a one-dimensional feature space would be insufficient since this one-dimensional manifold is cyclic, see Fig. 5, compare also with the  $2\pi$  period of an angle. Therefore, we have used a two-dimensional latent space. Further, only along the manifold in the latent space where the training data reside the decoder produces reasonable outputs. This can be further inspected by zooming in on a smaller region as displayed in Fig. 7.

To analyze the predictive performance of the two training methods, we define the fitting quality as

$$\text{FIT}_p = 1 - \sqrt{\frac{1}{NM} \sum_{t=1}^N \|y_t - \hat{y}_{t|t-p}\|^2}. \quad (12)$$

As a reference, the predictive performance is compared with a naive prediction using the previous frame at time step  $t - p$  as the prediction at  $t$  as  $\hat{y}_{t|t-p} = y_{t-p}$ . The result for a prediction horizon ranging from  $p = 0$  to  $p = 8$  is displayed in Fig. 6.

Clearly, joint learning (blue) outperforms separate learning in terms of predictive performance for prediction horizons greater than 0. Even by using the last available image frame for prediction (const. pred., brown), we obtain a better fit than the model that learns its parameter sequentially (red). This is due to the fact that the dynamical model often predicts frames, which do not correspond to any real pendulum, see Fig. 4, leading to a poor fit. Furthermore, joint training gives better predictions than the naive prediction. The predictive performance slightly degrades when the prediction horizon  $p$  increases, which is to be expected. Finally we also compare with the subspace identification method [18] (black, starred), which is restricted to linear models. Such a restriction does not capture the non-linear, embedded features and, hence, the predictive performance is sub-optimal.

## 4.2 Experiment 2: Tile moving in the plane

In this experiment we simulated 601 frames of a moving tile in a  $51 \times 51 = 2601$  pixels image. The control inputs are the increments in position in both of the two Cartesian directions. As in the previous experiment, the image sequence was reduced to  $\dim(y_t) = 50$  prior to the parameter learning using PCA. A four layer 50-

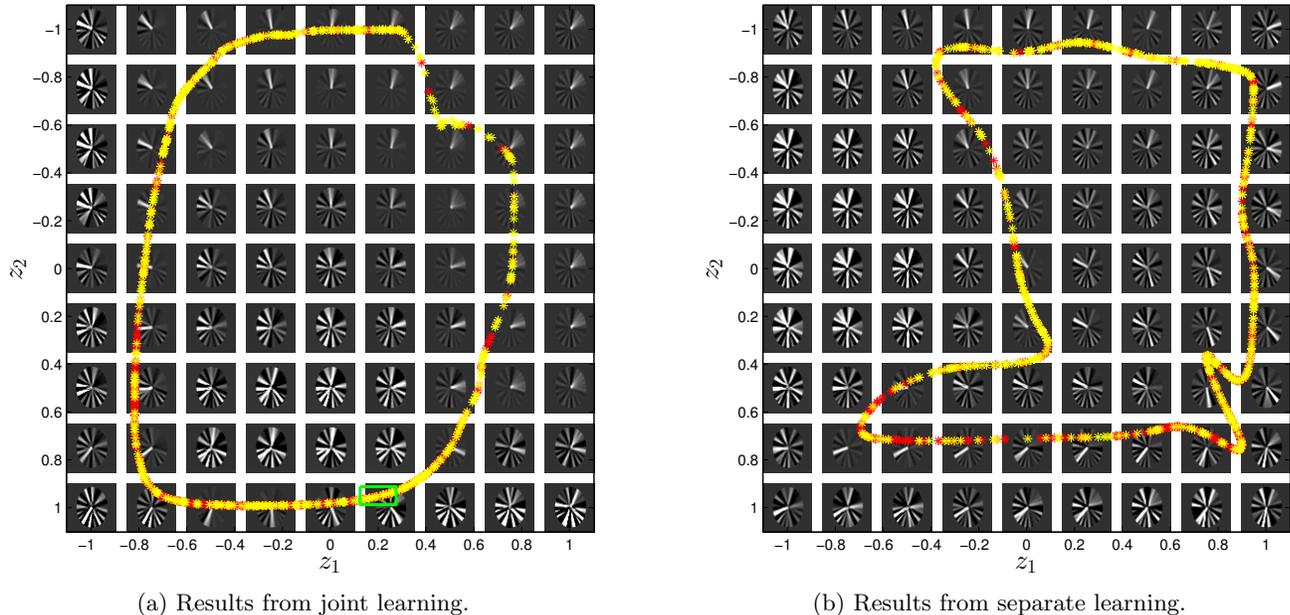


Figure 5: Exp. 1: The feature space  $z \in [-1, 1] \times [-1, 1]$  is divided into  $9 \times 9$  grid points. For each grid point the decoded high-dimensional image is displayed. The feature values corresponding to the training (red) and validation (yellow) data are displayed. Feature spaces found by joint (a) and separate (b) parameter learning. A zoomed in version of the green rectangle is presented in Fig. 7.

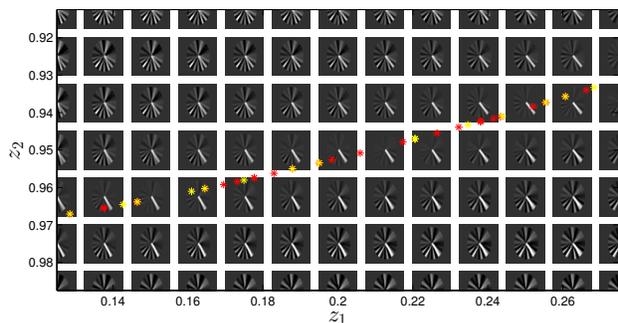


Figure 7: Exp. 1: A zoomed-in version of the feature space, corresponding to the green rectangle in Fig. 5a.

25-12-8-2 autoencoder was used for feature learning, and an 8-5-2 neural network for the dynamics.

As in the previous example, we evaluate the performance in terms of long-term predictions. The performance of joint training is illustrated in Fig. 8 on a validation data set. The model predicts future frames of the tile with high accuracy. In Fig. 9, the feature representation of the data is displayed. The features reside on a two-dimensional manifold encoding the two-dimensional position of the moving tile. The four corners in this manifold represent the four corners of the tile position within the image frame. This structure is induced by the dynamical description. The corresponding feature representation for the case of separate

learning does not exhibit such a structure, see the supplementary material.

In Fig. 10, the prediction performance is displayed, where our model achieves a substantially better fit than naively using the previous frame as prediction.

## 5 Discussion

From a system identification point of view, the prediction error method, where we minimize the one-step ahead prediction error, is fairly standard. However, in a future control or reinforcement learning setting, we are primarily interested in good prediction performance on a longer horizon in order to do planning. Thus, we have also investigated whether to additionally include a multi-step ahead prediction error in the cost (5). These models achieved similar performance, but no significantly better prediction error could be observed either for one-step ahead predictions or for longer prediction horizons.

Instead of computing the prediction errors in image space, see (5), we can compute errors in feature space to avoid a decoding step back to the high-dimensional space according to (4c). However, this will require an extra penalty term in order to avoid trivial solutions that map everything to zero, eventually resulting in a more complicated and less intuitive cost function.

Although joint learning aims at finding a feature rep-

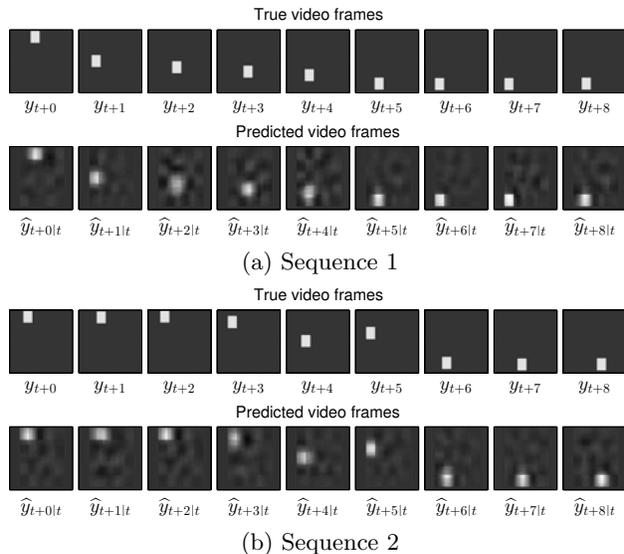


Figure 8: Exp. 2: True and predicted video frames on validation data.

resentation that is suitable for modeling the low-dimensional dynamical behavior, the pre-training initialization as described in Section 3.3 does not. If this pre-training yields feature values far from “useful” ones for modeling the dynamics, joint training might not find a good model.

The network structure has to be chosen before the actual training starts. Especially the dimension of the latent state and the order of the dynamics have to be chosen by the user, which requires a some prior knowledge about the system to be identified. In our examples, we chose the latent dimensionality based on insights about the true dynamics of the problem. In general, a model selection procedure will be preferable to find both a good network structure and a good latent dimensionality.

## 6 Conclusions and future work

We have presented an approach to non-linear system identification from high-dimensional time series data. Our model combines techniques from both the system identification and the machine learning community. In particular, we used a deep auto-encoder for finding low-dimensional features from high-dimensional data, and a nonlinear autoregressive exogenous model was used to describe the low-dimensional dynamics.

The framework has been applied to a pendulum moving in the horizontal plane. The proposed model exhibits good predictive performance and a major improvement has been identified by training the auto-encoder and the dynamical model jointly instead of

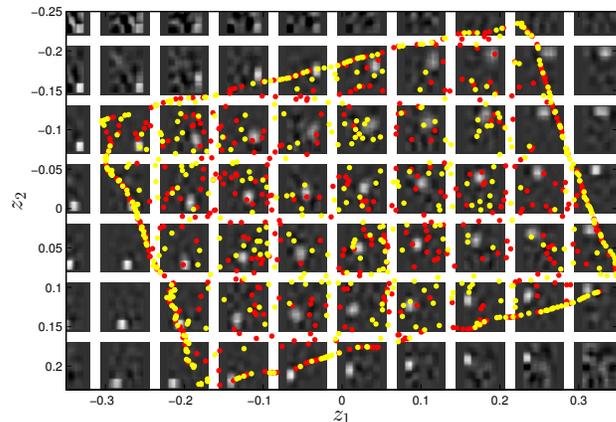


Figure 9: Exp. 2: Feature space zoomed in to the region  $z \in [-0.35, 0.35] \times [-0.25, 0.25]$ . The training (red) and validation (yellow) data reside on a two-dimensional manifold corresponding to the two-dimensional position of the tile.

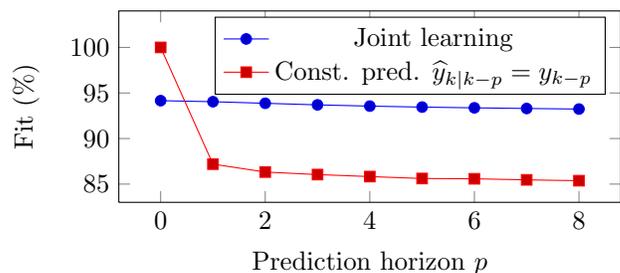


Figure 10: Experiment 2: The fit (12) for different prediction horizons  $p$ .

training them separately/sequentially.

Possible directions for future work include (a) robustify learning by using denoising autoencoders [19] to deal with noisy real-world data (b) apply convolutional neural networks, which are often more suitable for images; (c) exploiting the learned model for learning controller purely based on pixel information; (c) Sequential Monte Carlo methods will be investigated for systematic treatments of such nonlinear probabilistic models, which are required in a reinforcement learning setting.

In a setting where we make decisions based on (one-step or multiple-step ahead) predictions, such as optimal control or model-based reinforcement learning, a probabilistic model is often needed for robust decision making as we need to account for models errors [13, 3]. An extension of our present model to a probabilistic setting is non-trivial since random variables have to be transformed through the neural networks, and their exact probability density functions will be intractable to compute. Sampling-based approaches or determin-

istic approximate inference are two options that we will investigate in future.

### Acknowledgments

This work was supported by the Swedish Foundation for Strategic Research under the project *Cooperative Localization* and by the Swedish Research Council under the project *Probabilistic modeling of dynamical systems* (Contract number: 621-2013-5524). MPD was supported by an Imperial College Junior Research Fellowship.

### Supplementary material

In the second experiment in the paper, results the joint learning of prediction and reconstruction error have been reported. The joint learning brings structure to feature values, which is not present if the autoencoder is learn separately, see Fig. 11.

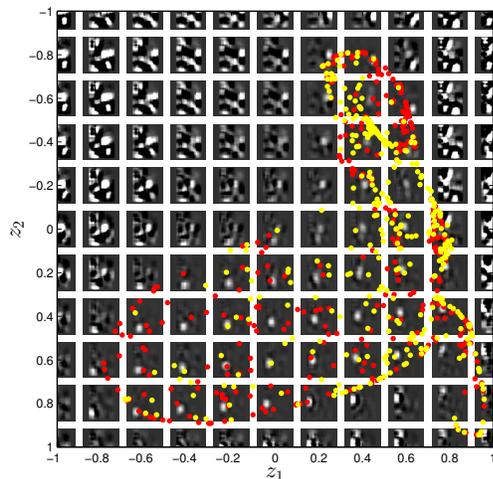


Figure 11: Exp. 2: The feature space in the region  $z \in [-1, 1] \times [-1, 1]$  for separate learning. The training (red) and validation (yellow) data does not reside on a structured a two-dimensional tile-formed manifold as it was the case for the joint learning in Fig. 9.

### References

- [1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
- [2] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294, 1988.
- [3] M. P. Deisenroth, D. Fox, and C. E. Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.
- [4] Z. Ghahramani. Learning dynamic bayesian networks. In C. Giles and M. Gori, editors, *Adaptive Processing of Sequences and Data Structures*, volume 1387 of *Lecture Notes in Computer Science*, pages 168–197. Springer, 1998.
- [5] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.
- [6] N. Kantas, A. Doucet, S. Singh, and J. Maciejowski. An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. In *Proceedings of the 15th IFAC Symposium on System Identification*, pages 774–785, Saint-Malo, France, July 2009.
- [7] N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, 2005.
- [8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [9] L. Ljung. *System identification, Theory for the user*. System sciences series. Prentice Hall, Upper Saddle River, NJ, USA, second edition, 1999.
- [10] L. Ljung. Perspectives on system identification. *Annual Reviews in Control*, 34(1):1–12, 2010.
- [11] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. New York, USA, 2 edition, 2006.
- [12] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
- [13] J. G. Schneider. Exploiting model uncertainty estimates for safe dynamic control learning. In *Advances in Neural Information Processing Systems (NIPS)*. Morgan Kaufman Publishers, 1997.
- [14] B. Schölkopf, A. J. Smola, and K.-R. Müller. Non-linear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998.
- [15] T. B. Schön, A. Wills, and B. Ninness. System identification of nonlinear state-space models. *Automatica*, 47(1):39–49, 2011.

- [16] R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.
- [17] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31(12):1691–1724, 1995.
- [18] P. Van Overschee and B. De Moor. *Subspace identification for linear systems - theory, implementation, applications*. Kluwer Academic Publishers, 1996.
- [19] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.