Extended Report 1–18

Linear Variational State-Space Filtering

Daniel Pfrommer DPFROM@SEAS.UPENN.EDU

Nikolai Matni NMATNI@SEAS.UPENN.EDU

Abstract

We introduce Variational State-Space Filters (VSSF), a new method for unsupervised learning, identification, and filtering of latent Markov state space models from raw pixels. We present a theoretically sound framework for latent state space inference under heterogeneous sensor configurations. The resulting model can integrate an arbitrary subset of the sensor measurements used during training, enabling the learning of semi-supervised state representations, thus enforcing that certain components of the learned latent state space to agree with interpretable measurements. From this framework we derive L-VSSF, an explicit instantiation of this model with linear latent dynamics and Gaussian distribution parameterizations. We experimentally demonstrate L-VSSF's ability to filter in latent space beyond the sequence length of the training dataset across several different test environments.

1. Introduction

Representation learning is central to many difficult machine learning problems. Uncovering low dimensional embeddings of high dimensional data enables novel reasoning about generative processes, data compression (Theis et al., 2017), and probabilistic forecasting (Ibrahim et al., 2021). Recent results in computer vision and natural language processing have demonstrated the effectiveness of large-scale generative models across difficult image (Yu et al., 2020; Zhang and Maire, 2020) and language (Devlin et al., 2018) tasks.

Contemporary generative representation learning techniques based on Variational Autoencoders (VAEs) (Kingma and Welling, 2014; Rezende et al., 2014) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are capable of generating high fidelity and realistic looking still images by leveraging powerful latent representations (van den Oord et al., 2017). Recent work using VAEs on video data and has shown the ability to predict future frames, (Babaeizadeh et al., 2017; Denton and Fergus, 2018) learn self-supervised representations of 3D structure (Lai et al., 2021), and enable compression ratios comparable to classical video codecs (Pessoa et al., 2020). In this paper we investigate the related problem of finding low dimensional embeddings suitable for control given pixel data and auxiliary low-dimensional (traditional) sensor measurements.

Applying deep representation learning techniques, specifically VAEs, to state space representations was investigated by Watter et al. (2015) and Krishnan et al. (2015), wherein a latent state embedding and locally linear latent state dynamics are jointly learned. More recent works such as Robust Controllable Embeddings (Banijamali et al., 2018), PlaNet (Hafner et al., 2019), Deep Variational Bayes Filters (Karl et al., 2017), Dream to Control (Hafner et al., 2020), Deep Kalman Smoother (DKS) (Krishnan et al., 2015), and Online Variational Filtering (OVF) (Campbell et al., 2021), have extended these techniques to more challenging systems with nonlinear dynamics and partial observability. These methods have several key limitations. With the exception of DKS and

OVF, these techniques either assume Markovian observation sequences (Watter et al., 2015; Banijamali et al., 2018) or use amortized posterior approximations, where smoothing variational posteriors are only partially conditioned on future measurements (Karl et al., 2017; Hafner et al., 2019; Lee et al., 2019; Hafner et al., 2020). In the latter case, these approximations have been shown by Bayer et al. (2021) to compromise the resulting the generative model, introducing a *conditioning gap* suboptimality, while methods assuming Markovian observation sequences must batch several images together in order to infer otherwise hidden state, such as velocity. In either case, these structural limitations mean these methods are inherently not capable of modelling the proper filtering posteriors.

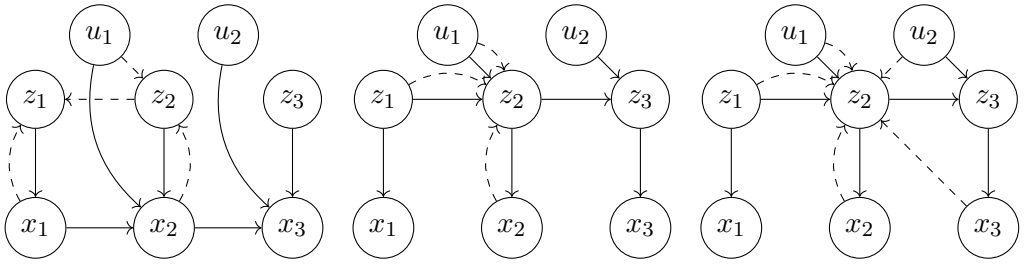
In this work we apply variational inference techniques to learn a dynamics-consistent model for filtering in a low dimensional latent state space. Since we additionally have access to the full smoothing distribution, we can both model hidden system state and do not have to partially condition our inference distribution. Paired with suitably powerful neural network architectures, we demonstrate the ability of our method to infer dynamics-consistent filtering models using only image data, as well as integrate information from different sensor modalities simultaneously. Although we only discuss inference on state spaces with linear dynamics in this paper, the method we present is applicable to any system where closed-form filtering priors and inverse dynamics models can be computed.

2. Related Work

Prior work in deep stochastic video generation and prediction methods such as (Babaeizadeh et al., 2017; Denton and Fergus, 2018; Lee et al., 2019), differ subtly from the state space learning problem we consider in this paper. Video generation methods seek to model $p(x_T|x_{1:T-1})$, where $x_{1:T}$ is a sequence of T video frames. By contrast, learned state space models are concerned with the inference distribution $p(z_{1:T}|x_{1:T}, u_{1:T-1})$ for some latent state $z_{1:T}$ and control input $u_{1:T-1}$. Whereas video generation models are free to pick an arbitrary latent state representation (e.g Babaeizadeh et al. (2017) use one latent state z across all frames $x_{1:T}$), state-space models are generally not free to do so.

Current deep variational state space learning methods can be broadly categorized into two classes: those that assume Markovian observation sequences (Watter et al., 2015; Banijamali et al., 2018), and those that allow for hidden state (Hafner et al., 2019; Karl et al., 2017). An observation sequence can generally be made Markovian by batching multiple image observations together, as done in (Watter et al., 2015; Banijamali et al., 2018). While this approach allows for models to only consider pairwise inference across temporally neighboring latent states z_t and z_{t+1}, it requires the mapping from latent state z_t to measurement x_t to be deterministic in order to preserve the Markovian assumption. This significant limitation has prompted work (Krishnan et al., 2015, 2016; Hafner et al., 2019; Bayer et al., 2021) into learning latent state spaces with both stochasticity and partial observability.

Learning latent state spaces with partial observability requires the challenging task of learning the entire conditional latent state trajectory distribution $p(z_{1:T}|x_{1:T}, u_{1:T-1})$. PlaNet (Hafner et al., 2019) and related models (Hafner et al., 2020; Lee et al., 2019) impose the factorization $p(z_{1:T}|x_{1:T}, u_{1:T-1}) = \prod_{t=1}^{T} p(z_t|z_{t-1}, x_t, u_{t-1})$. This effectively approximates the smoothing posterior $p(z_t|x_{1:T}, u_{1:T-1})$ by the filtering posterior $p(z_t|x_{1:t}, u_{1:t-1})$, incurring the conditioning gap suboptimality described in Bayer et al. (2021). The alternative approach taken by Krishnan et al.



(*a*) RCE Model (*b*) PlaNet Model (*c*) Our Model, DKS, OVF

Figure 1: Different latent model designs. Solid connections indicate the generative process, while dashed lines indicate the effective inference model for the future state z_2. Note (a) in Banijamali et al. (2018) (RCE Model), both z_1 and z_2 are inferred using a deterministic reverse transition (b) in Hafner et al. (2019) (PlaNet), there are no dependencies on future measurements x_3 or input u_2 when sampling z_2. Like Campbell et al. (2021) (OVF), our model (c) contains these dependencies but still enables access to the correct filtering distribution, unlike Krishnan et al. (2016) (DKS)

(2016) in DKS is to factor the distribution as $p(z_{1:T}|x_{1:T}, u_{1:T-1}) = \prod_{t=1}^{T} p(z_t|z_{t-1}, u_{t-1:T}, x_{t:T})$. However, DKS has the significant drawback that estimating z_t requires access to all future measurements $x_{t:T}$ as well as inputs $u_{t:T}$ and so does not provide access to a filtering distribution $p(z_t|x_{1:t})$. In contrast to these aforementioned works, we use a decomposition for the full smoothing posterior of a partially observable state space model that allows for variational inference using the proper smoothing distribution at training time, but that also enables access to a filtering distribution suitable for use in real-time systems. The same smoothing decomposition is proposed in the concurrent work Campbell et al. (2021), but we make the additional step of showing a closed form of the backwards smoothing distribution can be found, enabling simpler inference. The conceptual difference between these frameworks is illustrated in Fig. 1.

3. Problem Formulation

Consider the stochastic dynamics given by the following Markov state space model

$$z_1 \sim p_\theta(z_1), \; z_{t+1} \sim p_\psi(z_{t+1}|z_t, u_t) \tag{1}$$

with state z_t, input u_t, generative parameters θ and transition dynamics parameters ψ.

At each time step t, the system emits a set of k conditionally-independent observations $\mathcal{X}_t = \{x_t^{(j)}\}_{j=1}^{k}$ where

$$x_t^{(j)} \sim p_\theta(x_t^{(j)}|z_t). \tag{2}$$

The sensing modality can be different for each observation $x^{(j)}$, e.g., one observation could be an image obtained from a camera, and another a measurement obtained from an accelerometer. Further, in contrast to Watter et al. (2015) and related models, we do not require the latent state z_t to be inferrable from $\mathcal{X}_t$, i.e., our state space model is only partially observable.

Our framework also allows for arbitrary subsets of training sensing modalities to be deployed at test time. For example, suppose that training data is collected using both a motion capture system as well as images from a camera. Both the motion capture system and the camera data can be used during training, but the motion capture measurements can be omitted during testing.

Our goal is to learn both the parameters for the generative model θ as well as the system dynamics ψ. Formally, define $\mathcal{X}_{1:T} := \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T\}$ and $u_{1:T-1} := \{u_1, u_2, \ldots, u_{T-1}\}$ to be the T-length trajectories of measurements and control inputs, respectively. Given a set $\mathcal{D} = \{(\mathcal{X}^{(i)}_{1:T}, u^{(i)}_{1:T-1})\}_{i=1}^n$ of n independent input-observation trajectories of length T sampled from the true data-generating distribution $p_{\mathcal{D}}(\mathcal{D})$[1], we wish to find the maximum a posteriori estimate (MAP) for θ, ψ given by

$$\hat{\theta}, \hat{\psi} := \arg\max_{\theta,\psi} \; p(\theta, \psi|\mathcal{D}) = \arg\max_{\theta,\psi} \; p(\mathcal{D}_{\mathcal{X}}|\psi, \theta, \mathcal{D}_u),$$

under a uniform prior $p(\theta, \psi)$ on the model parameters, and where have let $\mathcal{D}_{\mathcal{X}} := \{\mathcal{X}^{(i)}_{1:T}\}_{i=1}^n$ and $\mathcal{D}_u := \{u^{(i)}_{1:T-1}\}_{i=1}^n$. Note that by independence of trajectories in the data set, the MAP estimate is equivalent to maximizing $\frac{1}{n}\sum_{i=1}^n \log p_{\theta,\psi}(\mathcal{X}^{(i)}_{1:T}|u^{(i)}_{1:T-1})$. To tackle this otherwise intractable optimization problem, we leverage variational inference techniques from Kingma and Welling (2014); Rezende et al. (2014) to introduce a variational approximation distribution $q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$ with inference parameters given by ϕ. Following Kingma and Welling (2014), the resulting evidence-based lower bound (ELBO) is

$$\begin{aligned}\mathcal{L}(\phi, \theta, \psi, \mathcal{X}_{1:T}, u_{1:T-1}) &= \log p_{\theta,\psi}(\mathcal{X}_{1:T}|u_{1:T-1}) - G \\ &= -\mathrm{D}_{\mathrm{KL}}(q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})||p_{\theta,\psi}(z_{1:T}|u_{1:T-1}))) \\ &\quad + \mathbb{E}_{q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})}[\log p_{\theta,\psi}(\mathcal{X}_{1:T}|z_{1:T})], \end{aligned} \tag{3}$$

where $G = \mathrm{D}_{\mathrm{KL}}(q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})||p_{\theta,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})))$ is the gap between the ELBO and the true log-likelihood of the generative conditional distribution.

Maximizing the ELBO is therefore equivalent to solving the original MAP parameter estimation problem under the following assumption.

Assumption 3.1 *The parametrization of the generating posterior $p_{\theta,\psi}(\mathcal{X}_{1:T}|z_{1:T}, u_{1:T-1})$ is sufficiently rich such that for optimal $\hat{\theta}, \hat{\psi}$ and true data-generating distribution $p_{\mathcal{D}}(\mathcal{X}_{1:T})$*

$$p_{\mathcal{D}}(\mathcal{X}_{1:T}) = p_{\hat{\theta},\hat{\psi}}(\mathcal{X}_{1:T}),$$

i.e., the true data-generating distribution can be feasibly modelled by our chosen parameterization. Additionally, the optimal variational posterior $q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$ is sufficiently rich such that

$$\operatorname*{ess\,sup}_{(\mathcal{X},u)\sim p_{\mathcal{D}}} \mathrm{D}_{\mathrm{KL}}(q_{\hat{\phi},\hat{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})||p_{\hat{\theta},\hat{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}))) = 0,$$

i.e., the gap between the lower bound and the true model log-likelihood can be driven to zero for the MAP estimate of θ, ψ.

These two conditions imply that (a) the inferred latent state distribution prior $q_{\hat{\phi}}(z_t) = \int q_{\hat{\phi}}(z_t|\mathcal{X}_t)p_{\mathcal{D}}(\mathcal{X}_t)d\mathcal{X}_t$ is equivalent to $p_{\hat{\theta}}(z_t)$ for all t, and (b) because the evidence lower bound is tight for the MAP parameter estimate, maximizing the lower bound is equivalent to solving the MAP optimization problem. While this is a very restrictive assumption, this is only used to show the correctness of subsequent problem simplifications in Section 4.1. Note that in practice it is sufficient for these conditions to be approximately satisfied in order to obtain good performing models. We leave the problem of quantifying the effect of approximately satisfying Assumption 3.1 on the resulting model degradation to future work.

1. We use the notation $p_{\mathcal{D}}(X)$ to denote the distribution of X as drawn from the true data generating distribution.

3.1. Smoothing Inference

We now propose a factorization of the full smoothing posterior $q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$ in terms of the observation inference model $q_\phi(z_t|\mathcal{X}_t^{(j)})$, state prior $p_\theta(z_t)$, and transition model $p_\psi(z_{t+1}|z_t, u_t)$, by leveraging the joint distribution smoothing decomposition used in Campbell et al. (2021); Briers et al. (2010). Rather than adopting a forward factorization of the smoothing posterior given by $q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$, wherein the current latent state z_t is dependent on the previous latent state z_{t-1}, we instead model the *backwards in time* dependency of z_t on z_{t+1} using the decomposition

$$q_{\phi,\psi}(z_t|z_{t+1}, \mathcal{X}_{1:t}, u_{1:t}) \propto q_\psi(z_{t+1}|z_t, u_t) q_{\phi,\psi}(z_t|\mathcal{X}_{1:t}, u_{1:t-1}). \tag{4}$$

Therefore given a latent state z_{t+1}, the distribution over the previous latent state z_t can be found by: (i) computing the filtering posterior $q_{\phi,\psi}(z_t|\mathcal{X}_{1:t}, u_{1:t-1})$, (ii) finding a closed form for the transition distribution $q_\psi(z_{t+1}|z_t, u_t) = p_\psi(z_{t+1}|z_t, u_t)$ as a function z_t, and (iii) taking the product of the densities. In particular, for linear dynamics and Gaussian distribution setting we consider in §4, all of these steps admit closed-form expressions. Furthermore because the full smoothing posterior factors as the product

$$q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}) = q_{\phi,\psi}(z_T|\mathcal{X}_{1:T}, u_{1:T-1}) \prod_{t=1}^{T-1} q_{\phi,\psi}(z_t|z_{t+1}, \mathcal{X}_{1:t}, u_{1:t}), \tag{5}$$

the samples and associated likelihoods for $q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$ can be efficiently computed recursively by first sampling the final latent state z_T from the marginal $q_{\phi,\psi}(z_T|\mathcal{X}_{1:T}, u_{1:T-1})$, and then recursively sampling from the corresponding marginals $q(z_t|z_{t+1}, \cdot)$ for $t = T-1, T-2, \ldots 1$.

Sampling from $q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$ is therefore a two-pass algorithm. We first perform a forward pass to compute the filtering posteriors $q_{\phi,\psi}(z_t|\mathcal{X}_{1:t}, u_{1:t-1})$ using the standard propagation and update recursions (Tanizaki, 1996)

$$q_{\phi,\psi}(z_t|\mathcal{X}_{1:t-1}, u_{1:t-1}) = \int q_\psi(z_t|z_{t-1}, u_{t-1}) q_\phi(z_{t-1}|\mathcal{X}_{1:t-1}, u_{1:t-2}) dz_{t-1}, \tag{6}$$

$$q_{\phi,\psi}(z_t|\mathcal{X}_{1:t}, u_{1:t-1}) \propto q_{\phi,\psi}(z_t, |\mathcal{X}_{1:t-1}, u_{1:t-1}) \prod_{j=1}^{k} \frac{q_\phi(z_t|x_t^{(j)})}{q_{\phi,\psi}(z_t)}. \tag{7}$$

Then we perform a backwards pass to sample a (backwards) trajectory $z_{T:1}$ using equations (4) and (5). This two-pass approach to sampling $z_{1:T}$ is visualized in Figure 2. Modelling the distribution $q_{\phi,\psi}(z_t)$ needed for the marginal posterior $q_{\phi,\psi}(z_t|\mathcal{X}_{1:t}, u_{1:t-1})$ is generally intractable, but Assumption 3.1 allows setting $q_{\phi,\psi}(z_t) = p_\theta(z_t)$ without affecting the correctness of the resulting model given proper parameterizations (see §A.3, as well as the conditions required for proper inference of per-observation $q_\phi(z_t|x_t^{(j)})$ discussed in §A.4).

With this sampling procedure, we can compute a Monte-carlo approximation for the KL term $\mathrm{D}_{\mathrm{KL}}(q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})||p_{\theta,\psi}(z_{1:T}|u_{1:T-1}))$ of the ELBO (3), as described in Kingma and Welling (2014), as the prior log-likelihood factors as

$$\log p_{\theta,\psi}(z_{1:T}|u_{1:T-1}) = \log p_\theta(z_1) + \sum_{t=1}^{T-1} \log p_\psi(z_{t+1}|z_t, u_t),$$

due to the Markov structure of the latent state. Similarly we can approximate the reconstruction term $\mathbb{E}_{q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})}[\log p_{\theta,\psi}(\mathcal{X}_{1:T}|z_{1:T})]$ of the ELBO (3) by leveraging conditional independence

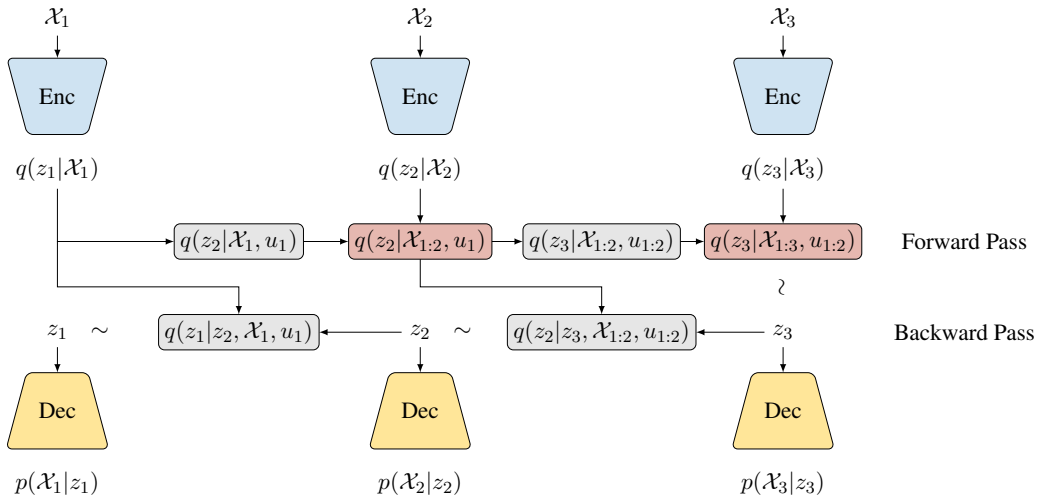


Figure 2: Visualization of the two-pass sampling approach for the smoothing posterior $q(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$ and corresponding reconstruction $p(\mathcal{X}_t|z_t)$. The red boxes show the observation-updated distributions and the grey boxes show the filtering-forwards/smoothing-backwards propagated distributions.

of the measurements $x_t^{(j)}$ given the latent state z_t:

$$\log p_{\theta,\psi}(\mathcal{X}_{1:T}|z_{1:T}) = \sum_{t=1}^{T}\sum_{j=1}^{k} \log p_{\theta,\psi}(x_t^{(j)}|z_t).$$

In the next section, we show that under suitable parameterizations of the dynamics $p_\psi(z_{t+1}|z_t, u_t)$, generative model $p_\theta(\mathcal{X}_t|z_t)$ and inference models $q_\phi(z_t|\mathcal{X}_t)$ we can efficiently maximize the ELBO (3). Detailed derivations of the results presented in this section con be found in Appendix A.

4. Variational State Space Models with Linear Dynamics

We consider an instantiation of the above framework with linear dynamics and Gaussian distributions. The general approach presented above is independent of these choices, and any combination of discrete or continuous latent variables with different distribution parameterizations can in principle be used for the latent state representation.

As a motivating example we consider the object tracking problem discussed in Section 3 where the latent state may be partially measured through a secondary sensor such as an accelerometer. In order to facilitate representation learning with partially known latent state, we will introduce two distinct observation models: a nonlinear model suitable for image data, as well as a linear model that can directly observe (subsets of) the latent state.

Consider a Markov state space model as described in Section 3, where the latent state $z_t \in \mathbb{R}^m$ and the system evolves according to linear dynamics of the form

$$z_1 \sim \mathcal{N}(0, \Sigma_z),\ z_{t+1} = Az_t + Bu_t + w_t,\ w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_w),$$

where the prior covariance Σ_z is fixed. The state space parameters (A, B, Σ_w) are treated either as known parameters or consolidated into the unknown parameters ψ, depending on the problem setup. We parameterize the measurement model $q_\phi(z_t|x_t^{(j)})$ as a Gaussian such that under Assumption 3.1

we have that

$$\frac{q_\phi(z_t|x_t^{(j)})}{q_{\phi,\psi}(z_t)} \approx \frac{q_\phi(z_t|x_t^{(j)})}{p_\theta(z_t)} \propto \mathcal{N}(h_t^{(j)}, H_t^{(j)}), \tag{8}$$

where $h_t^{(j)}$ and $H_t^{(j)}$ are computed from $x_t^{(j)}$ according to the model chosen for $p_\theta(x_t^{(j)}|z_t)$. We will discuss parameterizations of $h_t^{(j)}$ and $H_t^{(j)}$ for linear and nonlinear measurement models in §4.1.1 and §4.1.2.

Combining equation 8 with the linear dynamics model above, we find closed-form expressions for the mean and covariances of the resulting Gaussian filtering prior $q_{\phi,\psi}(z_t|\mathcal{X}_{1:t-1}, u_{1:t-1})$ and filtering posterior $q_{\phi,\psi}(z_t|\mathcal{X}_{1:t}, u_{1:t-1})$. Letting the filtering prior $q_{\phi,\psi}(z_t|\mathcal{X}_{1:t-1}, u_{1:t-1}) \sim \mathcal{N}(p_{t|t-1}, P_{t|t-1})$ and the filtering posterior $q_{\phi,\psi}(z_t|\mathcal{X}_{1:t}, u_{1:t-1}) \sim \mathcal{N}(p_{t|t}, P_{t|t})$, the priors can be computed recursively using the standard Kalman filter propagation equations (Terejanu, 2009):

$$P_{t|t-1} = AP_{t-1|t-1}A^\top + \Sigma_w, \quad p_{t|t-1} = Ap_{t-1|t-1} + Bu_t.$$

Similarly, the posterior $\mathcal{N}(p_{t|t}, P_{t|t})$ can be computed recursively in terms of the information matrix $P_{t|t}^{-1}$ and information vector $P_{t|t}^{-1}p_{t|t}$ using the Information Filter update equations (Terejanu, 2009):

$$P_{t|t}^{-1} = P_{t|t-1}^{-1} + \sum_{j=1}^{k}(H_t^{(j)})^{-1}, \quad P_{t|t}^{-1}p_{t|t} = P_{t|t-1}^{-1}p_{t|t-1} + \sum_{j=1}^{k}(H_t^{(j)})^{-1}h_t^{(j)}.$$

The use of the Information filter update as opposed to the standard state-space based update allows for easy simultaneous fusion of information from multiple observations. The reverse smoothing distribution $q_{\phi,\psi}(z_t|z_{t+1}, \mathcal{X}_{1:t}, u_{1:T-1}) = \mathcal{N}(\ell_t, L_t)$ can likewise be computed in terms of the filtering posterior $\mathcal{N}(P_{t|t}, p_{t|t})$ and next state z_{t+1}

$$L_t^{-1} = P_{t|t}^{-1} + A^\top\Sigma_w^{-1}A, \; L_t^{-1}\ell_t = A^\top\Sigma_w^{-1}(P_{t|t}^{-1} - Bu_t) + P_{t|t}^{-1}p_{t|t}.$$

We will now discuss different observation models for $\frac{q_\phi(z_t|x_t^{(j)})}{p_\theta(z_t)} \propto \mathcal{N}(h_t^{(j)}, H_t^{(j)})$ and $p_\theta(x_t^{(j)}|z_t)$ that can be used within this framework.

4.1. Observation Models

We make the following simplifying assumption.

Assumption 4.1 *The true underlying prior distribution $p_\mathcal{D}(z_t)$, state-conditional observation distribution $p_\mathcal{D}(\mathcal{X}_t|z_t)$, and state inference distributions $p_\mathcal{D}(z_t|\mathcal{X}_t)$ are time-invariant*

Using time-varying prior $p_\theta(z_t)$ would require the state inference distribution $q_\phi(z_t|\mathcal{X}_t)$ to also be time-varying. This is not only generally undesirable and computationally infeasible for large neural networks, but the resulting model would not be usable for sequences over a horizon longer than T, the horizon of the training data trajectories. Assumption 4.1 is satisfied, for example, if the initial state $p_\mathcal{D}(z_1)$ is equal to the steady-state distribution of the closed-loop system from which the data is generated and the observation model $p_\mathcal{D}(\mathcal{X}_t|z_t)$ is identical for all t.

4.1.1. Nonlinear Observation Model

To handle nonlinear and high-dimensional measurements $x_t^{(j)} \in \mathbb{R}^p$ such as raw pixel data, we introduce a nonlinear observation model. In this case we parameterize the generative model $p_\theta(x_t^{(j)}|z_t)$

for nonlinear sensing modality j as $p_\theta(x_t^{(j)}|z_t) = \mathcal{N}(\nu(x_t^{(j)}), \Sigma_x)$ for neural network-parameterized mean $\nu(x_t^{(j)})$ and fixed covariance matrix Σ_x, [2] and with corresponding inference model $q_\phi(z_t|x_t^{(j)}) = \mathcal{N}(r_h(x_t^{(j)}), r_H(x_t^{(j)}))$, for neural network-parameterized mean function r_h and positive definite neural network-parameterized coveriance matrix r_H. In Section 5 we discuss neural-network parameterizations of this model suitable for inference from image data, as well as how to parameterize r_H in order to enforce this constraint.

4.1.2. Linear Observation Model

Consider the standard linear measurement model where

$$x_t^{(j)} = C^{(j)} z_t + w_x,\ w_x \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_x),$$

where $C^{(j)} \in \mathbb{R}^{p \times n_z}$, and $p < n_z$, i.e., a noisy observation from a low-dimensional subspace of the latent-state is measured. In this case a closed form for both $p_\theta(x_t^{(j)}|z_t)$ and $p_\theta(z_t|x_t^{(j)})$ can be found, so the variational approximation $q_\phi(z_t|x_t^{(j)})$ is not needed, and the posterior $p_\theta(z_t|x_t^{(j)})$ can be used directly:

$$p_\theta(x_t^{(j)}|z_t) \sim \mathcal{N}(C^{(j)} z_t, \Sigma_x),\ p_\theta(z_t|x_t^{(j)}) \sim \mathcal{N}(\mu, \Sigma),$$

where $\Sigma^{-1} = C^\top \Sigma_x^{-1} C + \Sigma_z^{-1}$ and $\Sigma^{-1}\mu = C^\top \Sigma_x^{-1} x_t^{(j)}$.

Note that under Assumption 3.1 it follows that $(H^{(j)})^{-1} = C^\top \Sigma_x^{-1} C$ and $(H^{(j)})^{-1} h^{(j)} = C^\top \Sigma_x^{-1} x_t^{(j)}$, which correspond to the measurement information matrix/vector of a standard information filter (Terejanu, 2009).

By fixing the appropriate sparse $C^{(j)} \in \mathbb{R}^{s \times m}$ and combining the linear observation model with nonlinear observations, situations where the state is partially known or where a subset of the dataset is labelled can be cleanly handled.[3] We demonstrate this ability in the synthetic experiments presented in Section 5.

5. Experiments

We performed experiments in the following environments.

- **Pendulum Environment:** For the pendulum environment depicted in Figure 3, our dataset consists of 10000 sequences of length $T = 5$. We used a $3.14 \tanh(\theta/3.14)$ nonlinearity before rendering the final pendulum images in order to constrain the visual angle of the pendulum to the $(-\pi, \pi)$ range, even if the magnitude of θ sporadically exceeds π.
- **Blocks Environment:** For the second environment we moved a camera in the Unreal-Engine-based Airsim simulator (Shah et al., 2017) with double-integrator dynamics in x, y and fixed height z as well as fixed camera heading. The goal in this problem is to learn a mapping from

2. Current state of the art variational autoencoders such as VQ-VAE (van den Oord et al., 2017) and VDVAE (Child, 2020) use discrete logistic distribution mixtures parameterizations for $p_\theta(x_t^{(j)}|z_t)$. We use a diagonal normal distribution in our experiments since it is straightforward to implement and corresponds directly to a mean-square-error loss term in the resulting ELBO while still producing good results. We leave the problem of tuning this model for better performance to future work.
3. If $C^{(j)}$ is fixed, the prior $p_\theta(z)$ should be chosen to be consistent with the true distribution $p_\mathcal{D}(C^{(j)} z)$.

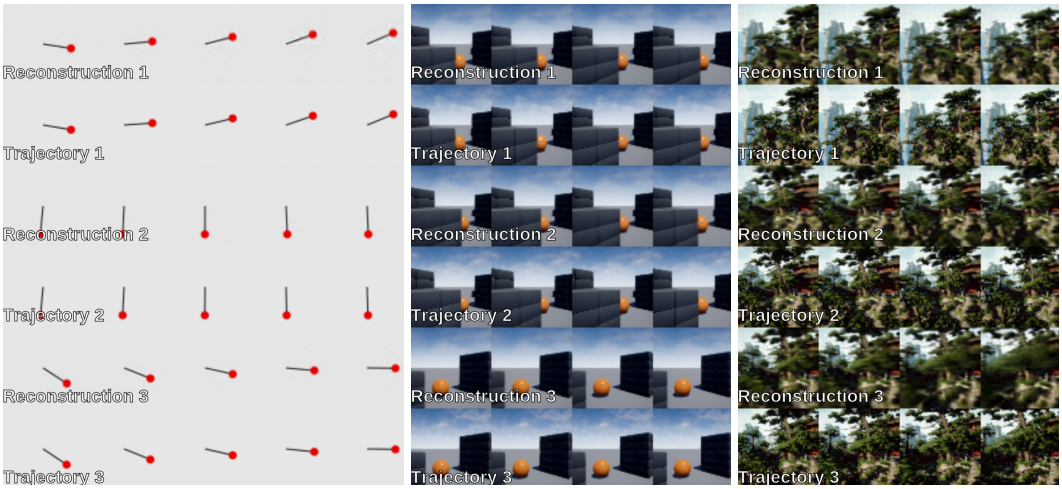


Figure 3: Reconstructions for three test trajectories from each of the pendulum, Blocks, and Zhangjiajie environments under fixed ψ. We speculate that the reconstructions for the Zhangjiajie environment are worse due to the feature-rich environment and dynamic fog effects present. The complex cloud patterns in the Blocks environment demonstrate a similar blurring effect.

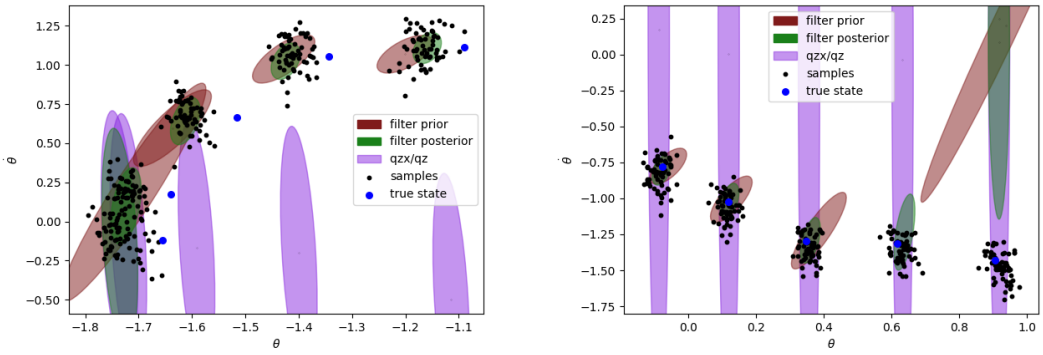
camera images to (x, y)-coordinates, a task analogous, but not equivalent to, visual-inertial-odometry (VIO). We used a dataset of size 10000 and trajectory length $T = 4$ in the "Blocks" Airsim environment.

- **Zhangjiajie Environment:** We additionally created a third dataset with the same double-integrator dynamics as the Blocks dataset, set in the "Zhangjiajie" Airsim environment. The Blocks and Zhangjiajie environments are both visualized in Fig 3. The Zhangjiajie environment is much more complex than the pendulum or block environments and we perform correspondingly worse on this dataset.

We primarily consider the case where the latent state dynamics parameters ψ are known, and seek to learn the model parameters θ, ϕ. We have found that simultaneously optimizing over ψ, θ, ϕ is difficult due to the tendency for the state transition matrix A to degenerate and cause posterior collapse. We hypothesize that prior work (Karl et al., 2017; Krishnan et al., 2016, 2015; Hafner et al., 2019, 2020) does not suffer from this issue since the estimates of the initial state z_1 are made independently of the dynamics parameters, whereas we model z_1 as dependent on $z_{2:T}$ and the system dynamics. In practice the collapse of $q_\phi(z_t)$ can be avoided through the introduction of a direct linear state observation model.

5.1. Qualitative Results

The trained $q_\phi(z_t|x_t)$ models and corresponding filter covariances visualized in Figure 4 for the pendulum environment demonstrate the ability to learn independence of the image observations and the hidden state consistent with a given dynamics model, as well as the ability to approximately recover the true underlying state simply by matching the prior $p_\theta(z_1)$. In Figure 3 we additionally show that the reconstructions of the samples from the smoothing distribution $q_\phi(z_{1:T}|x_{1:T}, u_{1:T})$ visualize the same dynamics as the sample image sequence.



(*a*) Image-based inference using pixel data only.

(*b*) Image-based inference with direct θ observations at training-time.

Figure 4: With fixed dynamics parameters ψ and setting $p_\theta(z_1) = p_\mathcal{D}(z_1)$, we demonstrate (a) the ability to learn a reasonable filter for $\theta, \dot{\theta}$ from image observations alone purely through matching the $p_\theta(z_1)$ prior and (b) higher accuracy filtering with additional linear measurement given by $x_t^{(2)} = \theta_t$ at training time

Supervision	None		Partial		Full	
	ψ fixed	ψ unknown	ψ fixed	ψ unknown	ψ fixed	ψ unknown
Pendulum	0.0458	0.558*	0.000452	0.000369	0.000211	0.000368
Blocks	0.244	9.715*	0.00292	0.00488	0.00193	0.00230
Zhangjiajie	2.308	6.186*	0.0519	0.137	0.0089	0.0413

Table 1: The mean squared position/angle error between the filtering posterior mean and the ground truth for different models on extended trajectories of length 200. For models where ψ is optimized, we use the learned ψ to evaluate the filter. (*) Note that these models are all degenerate.

5.2. Partial State Supervision

An advantage of our approach over prior work is the ability to use a subset of the training-time observations without compromising the correctness of the learned model. By introducing a direct linear measurement for a subset of the state components, we can perform partial state supervision. We consider the partially supervised setting with measurements $x_t^{(j)} = \theta$ and $x_t^{(j)} = \begin{bmatrix} x & y \end{bmatrix}$ for the pendulum and Airsim environments, respectively, as well as the fully supervised setting with measurements $x_t^{(j)} = \begin{bmatrix} \theta & \dot{\theta} \end{bmatrix}^\top$ and $x_t^{(j)} = \begin{bmatrix} x & y & \dot{x} & \dot{y} \end{bmatrix}^\top$ for the pendulum and Airsim environments, respectively. In all cases, we set the measurement noise covariance $\Sigma_x = 0.05I$. As shown in Table 1, using even partial supervision dramatically reduces the filtering error. This improvement is also reflected in the better latent space structures visualized in Figure 6, where the partially supervised models have a scale more similar to that of the true states (in this case a θ range of $(-\pi, \pi)$ and an x, y range of $(-4, 4)$ for the pendulum and Blocks environments respectively). The effect of the improved latent state structure is also evident in the more accurate filtering trajectories shown in 5.

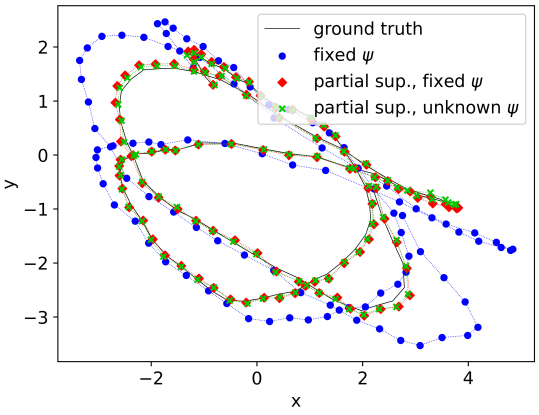


Figure 5: Filtering trajectories over $T = 100$ time steps in the Blocks environment for several different training configurations. Only pixels are used by the filter.

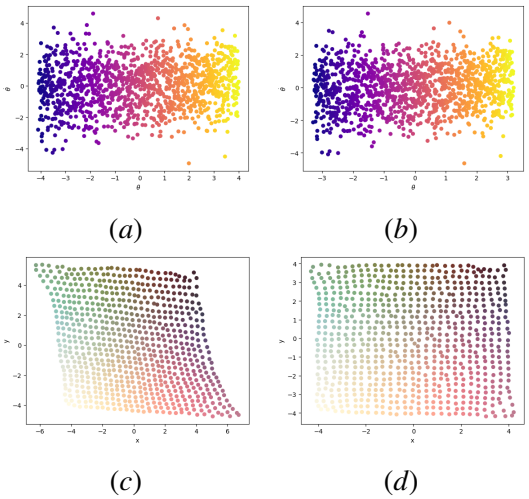


Figure 6: Visualizations of the latent space for the pendulum, (a) using only image data and (b) partially supervised, as well as the x, y of the Block environment, (c) from images and (d) partially supervised. Note that latent state spaces for partially supervised models are more consistent with the true scale in θ and x, y of $(-\pi, \pi)$ and $(-4, 4)$ respectively.

5.3. Extended Trajectory Experiments

To show that the resulting models learn a global state suitable for filtering on real-time systems, we considered the effect of evaluating the trained filtering model $q_\phi(z_t|\mathcal{X}_{1:t}, u_{1:t})$ on trajectories of length $T' = 200$ compared to our training time trajectory length of $T = 4$ and $T = 5$ for the Airsim and Pendulum environments respectively. Table 1 and Figure 5 demonstrate accurate filtering on the extended trajectories, even with unknown ψ given at least partial supervision. The minimal degredation under unknown ψ given supervision is notable as it suggests proper inference of the hidden state dynamics.

5.4. Implementation Details

All experiments were performed using Jax (Bradbury et al., 2018) in conjunction with Haiku (Hennigan et al., 2020) and Optax (Hessel et al., 2020). The code for generating the datasets and training the corresponding models is available online. [4]

For the nonlinear models described in §4.1.1, we used a modified version of the DCGAN (Radford et al., 2015) discriminator and generator architectures for the encoder and decoder networks respectively. For the encoder, we used an output dimensionality of 32 for the DCGAN discriminator and fed this into 3 x GELU (Hendrycks and Gimpel, 2020) activation + Linear layers (with 32 hidden variables) layers before reshaping the final output to be the correct shape for r_h. To ensure training stability of the resulting model under Assumption 3.1, $q_\phi(z_t|x_t^{(j)})/p_\theta(z_t)$ must be propor-

4. The codebase for these experiments is available at https://github.com/pfrommerd/variational_state_space_models

tional to a valid probability distribution. Therefore we must enforce that $r_H(x_t)^{-1} - \Sigma_z^{-1} \succeq 0$ for all x_t. In our implementation for simplicity we parameterize $r_H(x_t)^{-1}$ by the constant function $(L^\top L + \epsilon I)^{-1} + \Sigma_z^{-1}$ for a learned matrix $L \in \mathbb{R}^{n \times n}$ and $\epsilon = 0.0001$.

6. Conclusion

We introduced VSSF, a new family of VAE-based models that learn low dimensional state space representations and dynamics from high dimensional observation sequences, enabling real-time filtering over the latent state. Future work will look to extend the approach to beyond globally linear latent state space dynamics.

Acknowledgements

The authors thank Kostas Daniilidis for many helpful discussions as well as Bernadette Butcher for providing comments on the manuscript. NM is funded by NSF awards CPS-2038873, CAREER award ECCS-2045834, and a Google Research Scholar award. Work done by DP was partially funded by a Penn Grant for Faculty Mentoring Undergraduate Research (GFMUR).

References

Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.

Ershad Banijamali, Rui Shu, Hung Bui, Ali Ghodsi, et al. Robust locally-linear controllable embedding. In *International Conference on Artificial Intelligence and Statistics*, pages 1751–1759. PMLR, 2018.

Justin Bayer, Maximilian Soelch, Atanas Mirchev, Baris Kayalibay, and Patrick van der Smagt. Mind the gap when conditioning amortised inference in sequential latent-variable models, 2021.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. 2018. URL http://github.com/google/jax.

Mark Briers, Arnaud Doucet, and Simon Maskell. Smoothing algorithms for state–space models. *Annals of the Institute of Statistical Mathematics*, 62(1):61, 2010.

Paul Bromiley. Products and convolutions of gaussian probability density functions. *Tina-Vision Memo*, 3(4):1, 2003.

Andrew Campbell, Yuyang Shi, Thomas Rainforth, and Arnaud Doucet. Online variational filtering and parameter learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination, 2020.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2020.

Tom Hennigan, Trevor Cai, Tamara Norman, and Igor Babuschkin. Haiku: Sonnet for JAX. 2020. URL http://github.com/deepmind/dm-haiku.

Matteo Hessel, David Budden, Fabio Viola, Mihaela Rosca, Eren Sezener, and Tom Hennigan. Optax: composable gradient transformation and optimisation, in jax! 2020. URL http://github.com/deepmind/optax.

Mohamed R Ibrahim, James Haworth, Aldo Lipani, Nilufer Aslam, Tao Cheng, and Nicola Christie. Variational-lstm autoencoder to forecast the spread of coronavirus across the globe. *PloS one*, 16 (1):e0246120, 2021.

Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

Rahul G Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models, 2016.

Zihang Lai, Sifei Liu, Alexei A Efros, and Xiaolong Wang. Video autoencoder: self-supervised disentanglement of static 3d structure and motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9730–9740, 2021.

Alex X Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *arXiv preprint arXiv:1907.00953*, 2019.

Jorge Pessoa, Helena Aidos, Pedro Tomás, and Mário AT Figueiredo. End-to-end learning of video compression using spatio-temporal autoencoders. In *2020 IEEE Workshop on Signal Processing Systems (SiPS)*, pages 1–6. IEEE, 2020.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

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.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017. URL https://arxiv.org/abs/1705.05065.

Hisashi Tanizaki. *Nonlinear filters: estimation and applications*, volume 400. Springer Science & Business Media, 1996.

Gabriel Terejanu. Discrete kalman filter tutorial. 2009.

Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *CoRR*, abs/1711.00937, 2017. URL http://arxiv.org/abs/1711.00937.

Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. *arXiv preprint arXiv:1506.07365*, 2015.

Evan M. Yu, Adrian V. Dalca, Juan Eugenio Iglesias, and Mert R. Sabuncu. An auto-encoder strategy for adaptive image segmentation. In *Medical Imaging with Deep Learning*, 2020. URL https://openreview.net/forum?id=aEQCZR3xEm.

Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. *arXiv preprint arXiv:2012.03044*, 2020.

Appendix A. Supplementary Proofs

A.1. Evidence Lower Bound

We use the same general ELBO structure as Kingma and Welling (2014), adapted for our problem domain. For completeness we reproduce the derivation of the ELBO below.

Proposition 1 *The Evidence Lower Bound* $\mathcal{L}(\theta, \phi, \psi, \mathcal{X}_{1:T}, u_{1:T-1}) \leq \log p_{\theta,\psi}(\mathcal{X}_{1:T}|u_{1:T-1})$ *given by*

$$\mathcal{L}(\theta, \phi, \psi, \mathcal{X}_{1:T}, u_{1:T-1}) = \log p_{\theta,\psi}(\mathcal{X}_{1:T}|u_{1:T-1}) - G,$$

where $G = \mathrm{D}_{\mathrm{KL}}(q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})||p_{\theta,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}))$ *can be written*

$$\begin{aligned}\mathcal{L}(\theta, \phi, \psi, \mathcal{X}_{1:T}, u_{1:T-1}) = &-\mathrm{D}_{\mathrm{KL}}(q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})||p_{\theta,\psi}(z_{1:T}|u_{1:T-1})))\\ &+ \mathbb{E}_{q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})}[\log p_{\theta,\psi}(\mathcal{X}_{1:T}|z_{1:T})].\end{aligned}$$

Proof. Using $\mathbb{E}_{q_{\phi,\psi}}$ *as shorthand for* $\mathbb{E}_{q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})}$.

$$\begin{aligned}\mathcal{L}(\theta, \phi, \psi, \mathcal{X}_{1:T}, u_{1:T-1}) &= -\mathrm{D}_{\mathrm{KL}}(q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})||p_{\theta,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}))\\ &\quad + \log p_{\theta,\psi}(\mathcal{X}_{1:T}|u_{1:T-1})\\ &= \mathbb{E}_{q_{\phi,\psi}}[-\log q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}) + \log p_{\theta,\psi}(\mathcal{X}_{1:T}|u_{1:T-1})\\ &\quad + \log p_{\theta,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})]\\ &= \mathbb{E}_{q_{\phi,\psi}}[-\log q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}) + \log p_{\theta,\psi}(z_{1:T}, \mathcal{X}_{1:T}|u_{1:T-1})]\\ &= \mathbb{E}_{q_{\phi,\psi}}[-\log q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}) + \log p_{\theta,\psi}(z_{1:T}|u_{1:T-1})\\ &\quad + \log p_{\theta,\psi}(\mathcal{X}_{1:T}|z_{1:T-1})]\\ &= -\mathrm{D}_{\mathrm{KL}}(q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})||p_{\theta,\psi}(z_{1:T}|u_{1:T-1})] + \mathbb{E}_{q_{\phi,\psi}}[\log p_{\theta,\psi}(\mathcal{X}_{1:T}|z_{1:T-1})]\end{aligned}$$

A.2. Factorization of $q(z_t|z_{t+1}, \mathcal{X}_{1:t}, u_{1:t})$

The factorization of $q(z_t|z_{t+1}, \mathcal{X}_{1:t}, u_{1:t})$ follows from a straightforward application of Bayes' rule.

$$q(z_t|z_{t+1}, X_{1:t}, u_{1:t}) = \frac{q(z_{t+1}|z_t, \mathcal{X}_{1:t}, u_{1:t})q(z_t|\mathcal{X}_{1:t}, u_{1:t})}{q(z_{t+1}|\mathcal{X}_{1:t}, u_{1:t})}$$

Since $\int q(z_t|z_{t+1}, X_{1:t}, u_{1:t})dz_t = 1$, the denominator $q(z_{t+1}|\mathcal{X}_{1:t}, u_{1:t})$ is just a normalization constant. Therefore

$$q(z_t|z_{t+1}, X_{1:t}, u_{1:t}) \propto q(z_{t+1}|z_t, u_{1:t})q(z_t|\mathcal{X}_{1:t}, u_{1:t})$$

A.3. Proof of Correctness for Using the Modified Prior

Theorem 2 *Provided Assumption 3.1 holds, maximizing the ELBO (3) over* $p_{\mathcal{D}}(\mathcal{X}_{1:T}|u_{1:T-1})$ *under the modification that* $q_{\theta,\psi}(z_t) := p_{\theta,\psi}(z_t)$ *in the posterior update equation (7),*

$$q_{\phi,\psi}(z_t|\mathcal{X}_{1:t}, u_{1:t-1}) \propto q_{\phi,\psi}(z_t|\mathcal{X}_{1:t-1}, u_{1:t-1}) \prod_{j=1}^{k} \frac{q_\phi(z_t|\mathcal{X}_t^{(j)})}{q_{\phi,\psi}(z_t)},$$

is equivalent to maximizing the original ELBO over $p_{\mathcal{D}}(\mathcal{X}_{1:T}|u_{1:T-1})$.

Proof. *Let $\overline{q}_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$ denote the smoothing distribution $q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$ as described in §3 with $q_{\phi,\psi}(z_t)$ replaced by $p_{\theta,\psi}(z_t)$ in Equation (7) and $\overline{\mathcal{L}}(\phi, \theta, \psi, \mathcal{X}_{1:T}, u_{1:T-1})$ denote the resulting modified ELBO.*

First we show any optimal $\hat{\theta}, \hat{\psi}, \hat{\phi}$ that maximizes the true ELBO $\mathbb{E}_{\mathcal{X}_{1:T}, u_{1:T-1}} \mathcal{L}(\cdot, \mathcal{X}_{1:T}, u_{1:T-1})$ is an optimal solution for the modified ELBO $\mathbb{E}_{\mathcal{X}_{1:T}, u_{1:T-1}} \overline{\mathcal{L}}(\cdot, \mathcal{X}_{1:T}, u_{1:T-1})$. For notational brevity we will omit expectations over $\mathcal{X}_{1:T}, u_{1:T-1}$.

Note that under Assumption 3.1, for parameters $\hat{\phi}, \hat{\psi}$ maximizing the true ELBO, the latent priors of the inference and generative distributions are equivalent, i.e $q_{\hat{\phi},\hat{\psi}}(z_t) = p_{\theta,\psi}(z_t)$. Therefore for any optimal $\hat{\phi}, \hat{\psi}$ for the original ELBO, the modified and unmodified inference distributions are the same, i.e $\overline{q}_{\hat{\phi},\hat{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}) = q_{\hat{\phi},\hat{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$. Since the modified ELBO is given by

$$\begin{aligned}\overline{\mathcal{L}}(\phi, \theta, \psi, \mathcal{X}_{1:T}, u_{1:T-1}) = \log p_{\theta,\psi}(\mathcal{X}_{1:T}|u_{1:T-1}) \\ - \mathrm{D}_{\mathrm{KL}}(\overline{q}_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}) \,||\, p_{\theta,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})),\end{aligned}$$

under optimal $\hat{\phi}, \hat{\psi}, \hat{\theta}$ for $\mathcal{L}$, the modified ELBO becomes

$$\begin{aligned}\overline{\mathcal{L}}(\hat{\phi}, \hat{\theta}, \hat{\psi}, \mathcal{X}_{1:T}, u_{1:T-1}) = \log p_{\hat{\theta},\hat{\psi}}(\mathcal{X}_{1:T}|u_{1:T-1}) \\ - \mathrm{D}_{\mathrm{KL}}(q_{\hat{\phi},\hat{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}) \,||\, p_{\hat{\theta},\hat{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})).\end{aligned}$$

By Assumption 3.1 the KL divergence term is zero and $\hat{\theta}, \hat{\psi}$ maximize $\log p_{\hat{\theta},\hat{\psi}}(\mathcal{X}_{1:T}, u_{1:T-1})$, so $\hat{\phi}, \hat{\psi}, \hat{\theta}$ must also maximize $\overline{\mathcal{L}}$.

Now consider any $\overline{\theta}, \overline{\phi}, \overline{\psi}$ that maximize $\overline{\mathcal{L}}$. By the existence of $\hat{\theta}, \hat{\phi}, \hat{\psi}$, in order for $\overline{\theta}, \overline{\phi}, \overline{\psi}$ to maximize $\overline{\mathcal{L}}$, it must be that

$$\overline{\mathcal{L}}(\overline{\phi}, \overline{\theta}, \overline{\psi}, \mathcal{X}_{1:T}, u_{1:T-1}) \geq \overline{\mathcal{L}}(\hat{\phi}, \hat{\theta}, \hat{\psi}, \mathcal{X}_{1:T}, u_{1:T-1}) = \log p_{\hat{\theta},\hat{\psi}}(\mathcal{X}_{1:T}|u_{1:T-1}).$$

Note that $\hat{\theta}, \hat{\psi}$ maximize $p_{\theta,\psi}(\mathcal{X}_{1:T}|u_{1:T-1})$. Since the KL divergence is non-negative this implies that

$$\log p_{\overline{\theta},\overline{\psi}}(\mathcal{X}_{1:T}|u_{1:T-1}) = \log p_{\hat{\theta},\hat{\psi}}(\mathcal{X}_{1:T}|u_{1:T-1}),$$

and

$$\operatorname*{ess\,sup}_{\mathcal{X}_{1:T}, u_{1:T-1} \sim p_{\mathcal{D}}(\mathcal{D})} \mathrm{D}_{\mathrm{KL}}(\overline{q}_{\overline{\phi},\overline{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}) \,||\, p_{\overline{\theta},\overline{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})) = 0$$

Therefore the inferred latent state prior matches the generative latent state prior, i.e $\overline{q}_{\overline{\phi},\overline{\psi}}(z_t) = p_{\overline{\theta},\overline{\psi}}(z_t)$, and consequently $\overline{q}_{\overline{\phi},\overline{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}) = q_{\overline{\phi},\overline{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$, implying that $\overline{\theta}, \overline{\phi}, \overline{\psi}$ also maximize the true ELBO.

A.4. Correctness of learned filter $q(z_t|\mathcal{X}_{1:t}, u_{1:t-1})$ and inference model $q(z_t|x_t^{(j)})$

Assumption 3.1 guarantees that parameters $\hat{\theta}, \hat{\phi}, \hat{\psi}$ which maximize $\mathbb{E}_{\mathcal{X}_{1:T}, u_{1:T-1}} \mathcal{L}(\cdot, \mathcal{X}_{1:T}, u_{1:T-1})$ parameterize the proper smoothing inference distribution, i.e for all $\mathcal{X}_{1:T}, u_{1:T-1} \sim p_{\mathcal{D}}(\mathcal{X}_{1:T}, u_{1:T-1})$,

$$q_{\hat{\phi},\hat{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}) = p_{\hat{\theta},\hat{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}).$$

This is not necessarily sufficient to guarantee that either $q_{\hat{\phi},\hat{\psi}}(z_t|\mathcal{X}_{1:t}, u_{1:t-1}) = p_{\hat{\theta},\hat{\psi}}(z_t|\mathcal{X}_{1:t}, u_{1:t-1})$ or $q_{\hat{\phi},\hat{\psi}}(z_t|x_t^{(j)}) = p_{\hat{\theta}}(z_t|x_t^{(j)})$ everywhere.

In order for the inferred filtering distribution to satisfy $q_{\hat{\phi},\hat{\psi}}(z_t|\mathcal{X}_{1:t}, u_{1:t-1}) = p_{\hat{\theta},\hat{\psi}}(z_t|\mathcal{X}_{1:t}, u_{1:t-1})$, we must additionally have that the distribution $q_{\psi}(z_{t+1}|z_t, u_t)$ used in the time-reversed decompo-

sition (4) satisfies (i) the dynamics model used by the inference distribution $q_{\phi,\psi}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1})$ and the prior $p_{\theta,\psi}(z_{1:T}, u_{1:T-1})$ must be the same such that $q_\psi(z_{t+1}|z_t, u_t) = p_\psi(z_{t+1}|z_t, u_t)$ and (ii) $q_\psi(z_{t+1}|z_t, u_t)$ must be nonzero for all z_t, z_{t+1}, u_t. We satisfy (i) by explicitly setting $q_\psi(z_{t+1}|z_t, u_t) = p_\psi(z_{t+1}|z_t, u_t)$ in our decomposition given in §3 and note that (ii) is satisfied by the multivariate Gaussian parameterization for L-VSSF given in §4. Under both of these conditions

$$q_{\hat{\phi},\hat{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}) = p_{\hat{\theta},\hat{\psi}}(z_{1:T}|\mathcal{X}_{1:T}, u_{1:T-1}).$$

Using equality, under the reverse decomposition given by Equation (5)

$$\implies \quad q_{\hat{\phi},\hat{\psi}}(z_t|z_{t+1}, \mathcal{X}_{1:t}, u_{1:t}) = p_{\hat{\theta},\hat{\psi}}(z_t|z_{t+1}, \mathcal{X}_{1:t}, u_{1:t}) \quad \forall\, t \in [0, T]$$

$$\implies \quad q_{\hat{\psi}}(z_{t+1}|z_t, u_t) q_{\hat{\phi},\hat{\psi}}(z_t|\mathcal{X}_{1:t}, u_{1:t-1}) \propto p_{\hat{\psi}}(z_{t+1}|z_t, u_t) p_{\hat{\theta},\hat{\psi}}(z_t|\mathcal{X}_{1:t}, u_{1:t-1}) \quad \forall\, t \in [0, T].$$

Using conditions (i) and (ii) we can conclude that the filtering distributions must be identical for all t

$$\implies \quad q_{\hat{\phi},\hat{\psi}}(z_t|\mathcal{X}_{1:t}, u_{1:t-1}) = p_{\hat{\theta},\hat{\psi}}(z_t|\mathcal{X}_{1:t}, u_{1:t-1}) \quad \forall\, t \in [0, T]. \tag{9}$$

It remains to be shown that $q_{\hat{\phi}}(z_t|x_t^{(j)})$ satisfy $q_{\hat{\phi}}(z_t|x_t^{(j)}) = p_{\hat{\theta}}(z_t|x_t^{(j)})$. From Equations (7), (9) and Assumption 3.1 it follows that

$$\prod_{j=1}^{k} q_{\hat{\phi}}(z_t|x_t^{(j)}) = \prod_{j=1}^{k} p_{\hat{\theta}}(z_t|x_t^{(j)}) \tag{10}$$

Provided that (iii) $q_{\hat{\phi}}(z_t|x_t^{(j)})$ is always nonzero and (iv) for all $k' > 1$

$$\prod_{j=1, j\neq k'}^{k} q_{\hat{\phi}}(z_t|x_t^{(j)}) = \prod_{j=1, j\neq k'}^{k} p_{\hat{\theta}}(z_t|x_t^{(j)})$$

it follows directly that $q_{\hat{\phi}}(z_t|x_t^{(j)}) = p_{\hat{\theta}}(z_t|x_t^{(j)})$.

For L-VSSF, (iii) is satisfied by the choice of multivariate Gaussian distributions and (iv) is satisfied if $k-1$ of the observation models satisfy $q_\phi(z_t|x_t^{(j)}) = p_\theta(z_t|x_t^{(j)})$. This holds for the linear observation model given in §4.1.2 and (iv) holds for all experiments run in §5.

We note that it is also possible to satisfy (iv) by maximizing an objective function given by the ELBO for the full distribution $\log p(\mathcal{X}_{1:T}|u_{1:t-1})$, in addition to an ELBO for

$$\sum_{k'=2}^{k} \log p(\{\{x_t^{(j)}\}_{j=1, j\neq k'}^{k}\}_{t=1}^{T}|u_{1:t-1}),$$

i.e by inferring both the full filtering distribution, as well as each filter where the observation $j = k'$ has been removed. We leave experimental verification of this approach as a matter for future work.

A.5. Derivation of $q(z_t|z_{t+1}, X_{1:t}, u_{1:t})$ for L-VSSF

Theorem 3 *For the linear dynamics formulation in Section 4 we have* $q_{\varphi,\psi}(z_t|z_{t+1}, X_{1:t}, u_{1:t}) = \mathcal{N}(\ell_t, L_t)$ *where*

$$L_t^{-1} = P_{t|t}^{-1} + A^\top \Sigma_w^{-1} A$$
$$L_t^{-1}\ell_t = A^\top \Sigma_w^{-1}(P_{t|t}^{-1} - Bu_t) + P_{t|t}^{-1} p_{t|t}$$

Proof. From Equation (5) it follows

$$\begin{aligned} q_{\phi,\psi}(z_t|z_{t+1}, \mathcal{X}_{1:t}, u_{1:t}) &\propto q_\psi(z_{t+1}|z_t, u_t) q_{\phi,\psi}(z_t|\mathcal{X}_{1:t}, u_{1:t-1}) \\ &\propto \mathcal{N}(z_{t+1}, Az_t + Bu_t, \Sigma_w)\mathcal{N}(z_t, p_{t|t}, P_{t|t}) \end{aligned}$$

Where $\mathcal{N}(x, \mu, \Sigma)$ *denotes the multivariate normal density function with mean* μ *and covariance matrix* Σ *evaluated at* x*. Since*

$$\begin{aligned} \mathcal{N}(z_{t+1}, Az_t + Bu_t, \Sigma_w) &\propto e^{(z_{t+1}-(Az_t+Bu_t))^\top \Sigma_w^{-1}(z_{t+1}-(Az_t+Bu_t))} \\ &\propto e^{(A^{-1}z_{t+1}-z_t+A^{-1}Bu_t))^\top A^\top \Sigma_w^{-1} A(A^{-1}z_{t+1}-z_t+A^{-1}Bu_t)} \end{aligned}$$

Therefore $\mathcal{N}(z_{t+1}, Az_t + Bu_t, \Sigma_w) \propto \mathcal{N}(z_t, A^{-1}z_{t+1} - A^{-1}Bu_t, (A^{-1})^\top \Sigma_w A^{-1})$*. Consequently*

$$q_{\phi,\psi}(z_t|z_{t+1}, \mathcal{X}_{1:t}, u_{1:t}) \propto \mathcal{N}(z_t, A^{-1}z_{t+1} - A^{-1}Bu_t, A^\top \Sigma_w A)\mathcal{N}(z_t, p_{t|t}, P_{t|t})$$

For the product of two multivariate normal density functions $g(x) = \mathcal{N}(x, a, A)\mathcal{N}(x, b, B)$ *we can write* $g(x) \propto \mathcal{N}(x, c, C)$ *where* $C^{-1} = A^{-1} + B^{-1}$*,* $C^{-1}c = A^{-1}a + B^{-1}b$ *(Bromiley, 2003). It follows directly that*

$$q_{\phi,\psi}(z_t|z_{t+1}, \mathcal{X}_{1:t}, u_{1:t}) \propto \mathcal{N}(z_t, \ell_t, L_t)$$

Where ℓ_t, L_t *are defined as above. Because* $q_{\phi,\psi}(z_t|z_{t+1}, \mathcal{X}_{1:t}, u_{1:t})$ *is a distribution over* z_t*, this is in fact an equality, completing the proof.*