1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–15

Comparing Interpretable Inference Models for Videos of Physical Motion

Michael Pearce MICHAELPEARCE@GOOGLE.COM
Silvia Chiappa CSILVIA@GOOGLE.COM
Ulrich Paquet UPAQ@GOOGLE.COM
DeepMind, London

Abstract

We consider the problem of inferring the dynamics of a moving ball from sequences of images. We assume that the observations are generated from a low-dimensional latent linear Gaussian state-space model through a nonlinear mapping. We compare two variational approximations in a controlled environment.

1. Introduction

Inferring the dynamics of moving objects from pixel observations has been the subject of extensive study in recent literature. Deep recurrent neural networks have demonstrated impressive results on prediction and related tasks (Babaeizadeh et al., 2018; Chiappa et al., 2017; Denton and Birodkar, 2017; Finn et al., 2016; Oh et al., 2015; Srivastava et al., 2015; Sun et al., 2016), and probabilistic extensions have enabled to additionally learn interpretable internal representations and rich probabilistic reasoning capabilities (Archer et al., 2015; Fraccaro et al., 2016; Gao et al., 2016; Krishnan et al., 2017).

Fraccaro et al. (2017) showed that, by treating positions as auxiliary variables, the dynamics can be described in the low-dimensional space of positions and velocities, e.g. by using a state-space model representation of Newtonian laws. This approach used a variational autoencoder (VAE)-type approximation (Kingma and Welling, 2014; Rezende et al., 2014) which typically requires annealing of evidence lower bound (ELBO) terms to ensure that the dynamics are correctly accounted for during training.

In this paper, we evaluate whether a more explicit use of the state-space dynamics in the variational distribution yields better posterior approximations while also using a more relaxed annealing schedule when optimizing the ELBO. This work is in line with other recent attempts to overcome the issue of decoupling between the generative model and variational distribution (Lin et al., 2018; Rezende and Viola, 2018) in VAE-type approaches, which is particularly severe for time-series.

2. A Generative Model for Videos of Physical Motion

Our observations are experimentally controlled sequences of images $x_{1:T} \equiv x_1, \ldots, x_T$ representing the movement of a ball in the two-dimensional plane, as white pixels on a black background (Fig. 1a); see App. A.1 for details. As a generative model, we assume that each x_t is rendered from a noisy position $a_t \in \mathbb{R}^2$ using a neural network (NN) that returns a

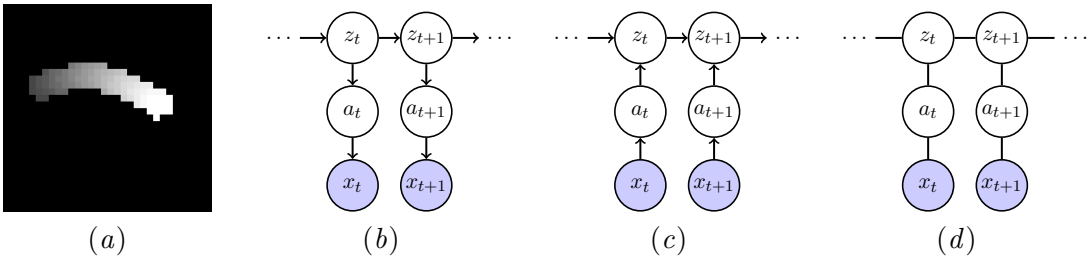

Figure 1: (a) An experimentally controlled video, with images overlaid over time. (b) The generative model in (2). (c) The directed graph variational approximation of (3). (d) The undirected graph variational approximation of (5).

Bernoulli probability for each pixel of a canvas, $p_\theta(x_t|a_t) = \mathcal{B}(x_t\,;\,\text{sigmoid}(\text{NN}(a_t;\theta)))$. We assume that a_t is generated from a linear Gaussian state-space model (LGSSM):

$$z_{t+1} = Az_t + u + \epsilon_t,\ \epsilon_t \sim \mathcal{N}(\epsilon_t; 0, \Sigma_z), \qquad a_t = Bz_t + \eta_t,\ \eta_t \sim \mathcal{N}(\eta_t; 0, \Sigma_a)\,, \tag{1}$$

with $z_1 = \mu + \epsilon_1$, $\epsilon_1 \sim \mathcal{N}(z_1; 0, \Sigma)$. We constrain the transition and emission matrices A and B to describe Newtonian dynamics, so that the hidden state $z_t \in \mathbb{R}^4$ represents the position and velocity, $a_t \in \mathbb{R}^2$ the noisy position, and $u \in \mathbb{R}^4$ allows modelling of a constant external force such as gravity. The joint distribution of all random variables factorizes as

$$p_{\theta,\gamma}(x_{1:T}, a_{1:T}, z_{1:T}) = \underbrace{\left\{\prod_{t=1}^T p_\theta(x_t|a_t)\right\}}_{\text{NN } p_\theta(x_{1:T}|a_{1:T})} \underbrace{p_\gamma(a_1|z_1)p_\gamma(z_1)\left\{\prod_{t=2}^T p_\gamma(a_t|z_t)p_\gamma(z_t|z_{t-1})\right\}}_{\text{LGSSM prior } p_\gamma(a_{1:T},z_{1:T})}, \tag{2}$$

where $\gamma = \{\mu, \Sigma, A, u, \Sigma_z, B, \Sigma_a\}$, $p_\gamma(z_t|z_{t-1}) = \mathcal{N}(z_t; Az_{t-1} + u, \Sigma_z)$, and $p_\gamma(a_t|z_t) = \mathcal{N}(a_t; Bz_t, \Sigma_a)$ (see Fig. 1b).

Due to the nonlinearity in $p_\theta(x_t|a_t)$, the marginal likelihood $p_{\theta,\gamma}(x_{1:T})$ and posterior distribution $p_{\theta,\gamma}(a_{1:T}, z_{1:T}|x_{1:T})$ are intractable. In Secs. 3 and 4, we introduce two different approximating distributions $q_{\phi,\gamma}(a_{1:T}, z_{1:T}|x_{1:T}) \approx p_{\theta,\gamma}(a_{1:T}, z_{1:T}|x_{1:T})$ for dealing with this problem (full details are given in App. B).

3. Directed Graph Variational Approximation

The LGSSM formulation of the latent dynamics allows us to approximate the intractable distribution $p_{\theta,\gamma}(a_{1:T}, z_{1:T}|x_{1:T})$ with the product of the tractable distribution $p_\gamma(z_{1:T}|a_{1:T})$ and an approximating distribution $q^{\text{D}}_\phi(a_{1:T}|x_{1:T}) = \prod_{t=1}^T q^{\text{D}}_\phi(a_t|x_t)$ for $p_{\theta,\gamma}(a_{1:T}|x_{1:T})$,

$$q^{\text{D}}_{\phi,\gamma}(a_{1:T}, z_{1:T}|x_{1:T}) = \left\{\prod_{t=1}^T q^{\text{D}}_\phi(a_t|x_t)\right\} p_\gamma(z_{1:T}|a_{1:T})\,. \tag{3}$$

Fig. 1c shows the approximation as a directed graphical model, which we distinguish with superscripts D. The factors $q^{\text{D}}_\phi(a_t|x_t) = \mathcal{N}(a_t\,;\,\mu_\phi(x_t),\,\Sigma_\phi(x_t))$ are diagonal Gaussians with shared neural network parameters ϕ. The approximation in (3) enables us to write the evidence lower bound (ELBO) $\mathcal{L}^{\text{D}}$ to $\log p_\theta(x_{1:T})$ as

$$\mathcal{L}^{\text{D}}(\theta,\gamma,\phi) = \mathbb{E}_{q^{\text{D}}_\phi}\left[\textstyle\sum_t \log p_\theta(x_t|a_t)\right] \underbrace{- \mathbb{E}_{q^{\text{D}}_\phi}\left[\textstyle\sum_t \log q^{\text{D}}_\phi(a_t|x_t)\right] + \mathbb{E}_{q^{\text{D}}_\phi}\left[\log p_\gamma(a_{1:T})\right]}_{-\text{KL}(q^{\text{D}}_\phi(a_{1:T}|x_{1:T})\,\|\,p_\gamma(a_{1:T}))}, \tag{4}$$

where expectations are over $q^{\mathrm{D}}_{\phi}(a_{1:T}|x_{1:T})$. Notice that γ does *not* appear in the distribution over which expectations are computed. The first and last terms are computed using Monte-Carlo while the middle term is computed analytically. This approach is similar to the one in Fraccaro et al. (2017), but differs as the ELBO averages over $q^{\mathrm{D}}_{\phi}(a_{1:T}|x_{1:T})$ rather than over $q^{\mathrm{D}}_{\phi}(a_{1:T}, z_{1:T}|x_{1:T})$. The Kullback-Leibler (KL) term in (4) indicates how close the positions that were estimated from the image sequence are to that of the LGSSM.

As the approximating distribution $q^{\mathrm{D}}_{\phi}(a_{1:T}|x_{1:T})$ does not explicitly incorporate the LGSSM distribution, we found that, to gives satisfactory results, it needs to be forced to "listen" to the LGSSM through a carefully annealed high to low weighting of the KL term. To alleviate this problem, in the next section we describe an alternative approach which makes more explicit use of the LGSSM dynamics in the variational distribution.

4. Undirected Graph Variational Approximation

We enable the conditional distribution $q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T})$ to incorporate or *marginalize* over plausible latent paths given by the LGSSM by allowing $q^{*}_{\phi}(a_t|x_t)$ to model uncertain "inputs" to the LGSSM, as Gaussian factors in an *undirected* (U) graphical model,

$$q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}, z_{1:T}|x_{1:T}) = \frac{1}{Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})} \left\{\prod_{t=1}^{T} q^{*}_{\phi}(a_t|x_t)\right\} p_{\gamma}(a_{1:T}|z_{1:T})\, p_{\gamma}(z_{1:T})\,. \tag{5}$$

In comparison to (3), we note that two distributions $q^{*}_{\phi}(a_t|x_t)$ and $p_{\gamma}(a_t|z_t)$ are multiplied together to yield a local belief for a_t. As is common to undirected graphical models, an additional normalizing constant $Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T}) = \int_{a_{1:T}} \int_{z_{1:T}} q^{*}_{\phi}(a_{1:T}|x_{1:T})\, p_{\gamma}(a_{1:T}|z_{1:T})\, p_{\gamma}(z_{1:T})$ is required. The undirected approximation is shown in Fig. 1d. This approximation is similar to the linear dynamical system approximation in Lin et al. (2018), but differs in the inclusion of the auxiliary variables $a_{1:T}$.

We write the ELBO as an average over $q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T})$[1]:

$$\mathcal{L}^{\mathrm{U}}(\theta,\gamma,\phi) = \mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}}\left[\textstyle\sum_t \log p_{\theta,\gamma}(x_t|z_t)\right] \underbrace{-\mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}}\left[\textstyle\sum_t \log \mathcal{N}(\mu^{*}_{\phi}(x_t); Bz_t, \Sigma_a + \Sigma^{*}_{\phi}(x_t))\right] + \log Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})}_{-\mathrm{KL}(q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T}) \,\|\, p_{\gamma}(z_{1:T}))}\,. \tag{6}$$

The expectations over $q^{\mathrm{U}}_{\phi,\gamma}(z_t|x_{1:T})$ are found by integrating out $a_{1:T}$ and using Kalman filtering and Rauch-Tung-Striebel smoothing over $z_{1:T}$ (Barber et al., 2011). The middle term is the cross entropy of Gaussians which can be computed in closed form. More details are given in App. B.2.

By modelling $a_{1:T}|x_{1:T}$ jointly as an average over latent $z_{1:T}$ trajectories, (5) gives substantially better ELBOs than (3), which uses a factorized inference "layer" for $a_t|x_t$. We illustrate this next experimentally in Sec. 5. Numerous other results are presented in App. C.

1. We might equally write an ELBO by averaging over $q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T})$. This idea is revisited in App. B.2.2.

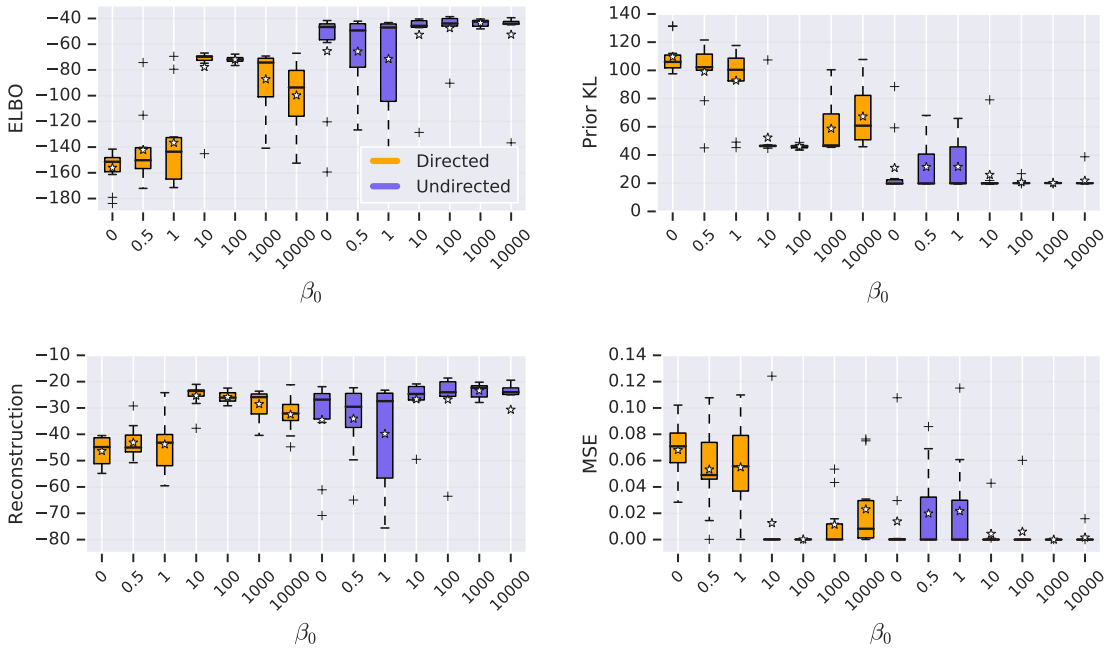

Figure 2: Comparison of the inference models described in Secs. 3 and 4 for different $\beta_0 \to \beta = 1$ KL annealing schedules. *Top left.* At the best local maximum, ELBO (5) is a better bound than (3) by around 28 nats over 30 frames. *Top right.* The KL terms of (3) and (5). *Bottom left.* The "reconstruction" terms of (3) and (5). *Bottom right.* Ground truth recovery of latent dynamics. The mean squared error (MSE) is relative to the ground truth (and scale invariant with respect to $q(a_{1:T}|x_{1:T})$). Over various seeds, $q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T})$ recovers the ground truth latent dynamics much more robustly and consistently than the factorized $q^{\mathrm{D}}_{\phi}(a_{1:T}|x_{1:T})$. Samples from both these distributions were matched and compared to ground truth trajectories; see App. A.3.

5. Experiments

We compared the inference models described in Secs. 3 and 4 on a synthesized dataset of videos of a cannonball fired with a random speed and angle (see App. A.1 for details). We multiplied the KL-term in the ELBO with a β-term, initialized at β_0, and annealed down to reach $\beta = 1$ at 10,000 gradient descent iterations (or annealed up, if $\beta_0 < 1$). After that, β was kept at one, so that the results highlight various local maxima which are not escaped from.

Fig. 2 shows the results for different $\beta_0 \to \beta = 1$ annealing schedules after 100,0000 training iterations (each experiment was repeated 10 times with different random initializations). We can see that the undirected inference model gives a substantially lower KL divergence from the prior dynamics with better or comparable reconstruction error, and that it is more robust to the annealing schedule.

References

E. Archer, I. M. Park, L. Buesing, J. Cunningham, and L. Paninski. Black box variational inference for state space models. *arXiv:1511.07367*, 2015.

M. Babaeizadeh, C. Finn, D. Erhan, R. Campbell, and S. Levine. Stochastic variational video prediction. In *6th International Conference on Learning Representations*, pages 1–14, 2018.

D. Barber, A. T. Cemgil, and S. Chiappa. *Bayesian Time Series Models*, chapter Inference and estimation in probabilistic time series models, pages 1–31. Cambridge University Press, 2011.

S. Chiappa, S. Racanière, D. Wierstra, and S. Mohamed. Recurrent environment simulators. In *5th International Conference on Learning Representations*, pages 1–61, 2017.

E. L. Denton and V. Birodkar. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems 30*, pages 4414–4423, 2017.

C. Finn, I. J. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems 29*, pages 64–72, 2016.

M. Fraccaro, S. K. Sønderby, U. Paquet, and O. Winther. Sequential neural models with stochastic layers. In *Advances in Neural Information Processing Systems 29*, pages 2199–2207, 2016.

M. Fraccaro, S. Kamronn, U. Paquet, and O. Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Advances in Neural Information Processing Systems 30*, pages 3604–3613, 2017.

Y. Gao, E. W. Archer, L. Paninski, and J. P. Cunningham. Linear dynamical neural population models through nonlinear embeddings. In *Advances in Neural Information Processing Systems 29*, pages 163–171, 2016.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations*, 2014.

R. Krishnan, U. Shalit, and D. Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2101–2109, 2017.

W. Lin, N. Hubacher, and M. E. Khan. Variational message passing with structured inference networks. In *6th International Conference on Learning Representations*, 2018.

J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh. Action-conditional video prediction using deep networks in Atari games. In *Advances in Neural Information Processing Systems 28*, pages 2863–2871, 2015.

D. J. Rezende and F. Viola. Taming VAEs. *arXiv:1810.00597*, 2018.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286, 2014.

N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 843–852, 2015.

W. Sun, A. Venkatraman, B. Boots, and J. A. Bagnell. Learning to filter with predictive state inference machines. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1197–1205, 2016.

Appendix A. Experimental Setup

A.1. Dataset

The dataset used for the experiments was generated using the following LGSSM representation of Newtonian laws

$$A = \begin{pmatrix} I & \delta I \\ 0 & I \end{pmatrix}, \qquad u = -g(0 \;\; 0.5\delta^2 \;\; 0 \;\; \delta)^{\mathsf{T}}, \qquad B = (I \;\; 0), \tag{7}$$

where I is the 2×2 identity matrix, $\delta = 0.015$ is the sampling period, and $g = 9.81$ is the gravitational constant. We used $\Sigma_z = 0$, and $\Sigma_a = 0.001I$. Each ball was shot with random shooting angle in the interval $(20°, 70°)$ from the left side of the x-axis in the interval $(-0.5, -0.1)$. The initial position on the y-axis was sampled in the interval $(-0.5, 0.5)$. The initial velocity was sampled in the interval $(2, 4)$.

To render the positions into white patches of radius $R = 2$ in the image, $a_{1:T}$ were re-scaled to the interval $[R, H-1-R] \times [R, W-1-R]$ where $H = 32$, $W = 32$ are the height and width of the image respectively. This re-scaling ensured that each ball was always fully contained in the image.

A.2. Network Architectures and Training

Both the encoder and decoder networks were fully connected networks with a single hidden layer of 1024 nodes. For the decoder network $p_\theta(x_t|a_t)$ the initial layer had two nodes for a_t and the final layer had 1024 nodes whose outputs were passed through the sigmoid function to return a Bernoulli probability for each pixel of the 32×32 canvas, x_t,

$$\begin{aligned} x_t \sim p_\theta(x_t|a_t) &= \mathcal{B}(x_t\,;\, \text{sigmoid}(\text{NN}(a_t;\theta)))\,, \\ \text{NN}(a_t;\theta) &= W_2^\theta \tanh(W_1^\theta a_t + b_1^\theta) + b_2^\theta\,. \end{aligned}$$

The learned parameters were thus all weight and bias matrices $\theta = \{W_1^\theta, W_2^\theta, b_1^\theta, b_2^\theta\}$.

The encoder networks, $q_\phi^D(a_t|x_t)$ and $q_\phi^*(a_t|x_t)$, had the same architecture with initial and final layers of 1024 and 4 nodes respectively, the final nodes outputting the mean and log variance of the approximate posterior,

$$h = \tanh(W_1^\phi x_t + b_1^\phi), \quad \mu_\phi(h) = W_\mu^\phi h + b_\mu^\phi, \quad \sigma_\phi(h) = \exp(W_\sigma^\phi h + b_\sigma^\phi)\,,$$

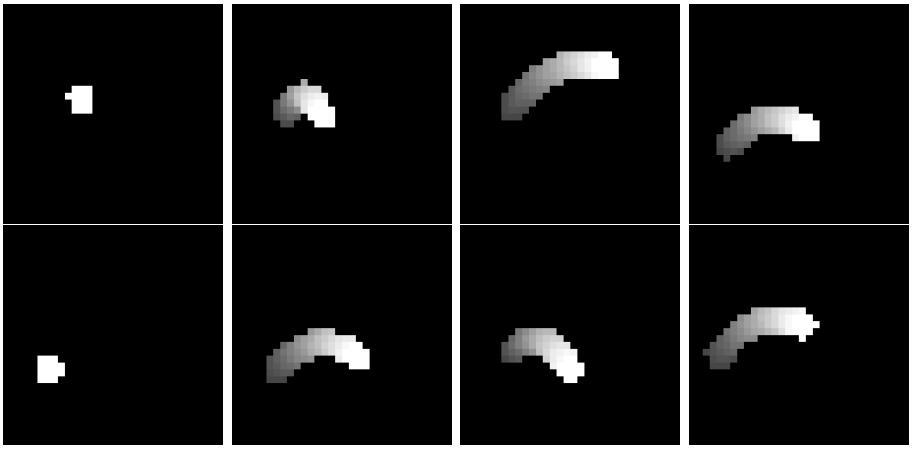

Figure 3: Example training data. *Left:* individual frames. *All others:* 30 frames superimposed with time index indicated by shade.

where $\mathcal{N}(a_t; \mu_\phi(h), \sigma^2_\phi(h))$ is the factor in the respective inference models; and the learned network parameters were $\phi = \{W^\phi_1, b^\phi_1, W^\phi_\mu, b^\phi_\mu, W^\phi_\sigma, b^\phi_\sigma\}$. All weight matrices were randomly initialized from $\mathcal{N}(\cdot; 0, 1/d)$, where d is the number of matrix columns, and all biases were initialized to zero.

We constrained the transition and emission matrices A and B as in (7), with δ initialized to 0.015, except for Fig. 6, where we instead simply initialized A to the 4×4 identity matrix and B to a 2×4 random matrix with elements sampled from $\mathcal{N}(\cdot; 0, 1)$. Σ, Σ_z and Σ_a were constrained to be diagonal and initialized to identity matrices, and μ was initialized to a random vector with elements sampled from $\mathcal{N}(\cdot; 0, 1)$.

The KL divergence term of the ELBO was exponentially annealed using the following schedule

$$\beta_i = \begin{cases} 1 + (\beta_0 - 1)\exp\left(-i/2000\right) & i \leq 10,000 \\ 1 & \text{otherwise} \end{cases}$$

(at iteration $i = 0$, we have $\beta = \beta_0$; at iteration $i = 10,000$, we have $\beta = 1 + (\beta_0 - 1)\mathrm{e}^{-5}$). We performed ablation studies without annealing, in which case $\beta_i = \beta_0$ until 10,000 iterations and $\beta_i = 1$ thereafter.

All parameters were optimized end-to-end using the Adam optimizer with a learning rate of 0.001 and default values $\beta_1 = 0.1$, $\beta_2 = 0.999$ (these two β's being optimizer parameters, not a KL annealing scalar) and $\epsilon = 10^{-8}$ with a minibatch size of 20 videos. Training was stopped after 100,000 iterations.

With these settings, one training iteration typically took 150ms for the directed model and 400ms for the undirected model on a Nvidia P100 GPU. We believe that this is due to the extra Cholesky decompositions that are required.

A.3. Ground Truth Mean Squared Error Computation

We measured the ability of the directed and undirected inference models to learn a physically plausible latent domain by using a linear model to predict the ground truth trajectory of the ball, $a^{\text{gt}}_{1:T}$, using a sample from the approximate posterior $\widehat{a}_{1:T}$.

For the directed model, we sampled the inference network directly, $\widehat{a}_{1:T} \sim \prod_{t=1}^{T} q^{\text{D}}_{\phi}(a_t|x_t)$. For the undirected model, we approximated samples from $q^{\text{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T})$ by computing the marginals $q^{\text{U}}_{\phi,\gamma}(z_t|x_{1:T}) = \mathcal{N}(z_t\,;\,\mu_{z_t},\,\Sigma_{z_t})$ with Kalman filtering and Rauch-Tung-Striebel smoothing (with a_t averaged out). We then drew samples $\widehat{a}_{1:T}$ from the average of $p_\gamma(a_t|z_t)$ over this marginal: $\widehat{a}_{1:T} \sim \prod_{t=1}^{T} \mathcal{N}(a_t; B\mu_{z_t}, B^{\mathsf{T}}\Sigma_{z_t}B + \Sigma_a)$.[2]

Using globally estimated parameters $W^{\text{MSE}} \in \mathbb{R}^{2\times 2}$ and $b^{\text{MSE}} \in \mathbb{R}^2$ to project (rotate, scale, move) each model's inferred trajectory onto the ground truth space. The mean squared error in the ground truth domain is given by

$$\text{MSE}(\widehat{a}_{1:T}; a^{\text{gt}}_{1:T}) = \frac{1}{T}\sum_t \left\| W^{\text{MSE}}\widehat{a}_t + b^{\text{MSE}} - a^{\text{gt}}_t \right\|^2 . \tag{8}$$

Equation 8 is the squared error from a linear regression model using $\widehat{a}_{1:T}$ to predict the ground truth $a^{\text{gt}}_{1:T}$. The parameters were estimated from N videos $\{x^n_{1:T}\}_{n=1}^N$ for which we had the ground truth $\{a^{\text{gt},n}_{1:T}\}$,

$$W^{\text{MSE}}, b^{\text{MSE}} = \arg\min_{W,b} \sum_n \sum_t \left\| W\widehat{a}^n_t + b - a^{\text{gt},n}_t \right\|^2 , \tag{9}$$

for each of the two models.

Appendix B. Approximations

The true posterior distribution from the generative model is given by

$$\begin{aligned} p_{\theta,\gamma}(a_{1:T}, z_{1:T}|x_{1:T}) &= \frac{p_\theta(x_{1:T}|a_{1:T})\, p_\gamma(a_{1:T}, z_{1:T})}{p_{\theta,\gamma}(x_{1:T})} \\ &= \frac{p_\theta(x_{1:T}|a_{1:T})\, p_\gamma(a_{1:T})}{p_{\theta,\gamma}(x_{1:T})} \, \frac{p_\gamma(a_{1:T}, z_{1:T})}{p_\gamma(a_{1:T})} \\ &= p_{\theta,\gamma}(a_{1:T}|x_{1:T})\, p_\gamma(z_{1:T}|a_{1:T})\,. \end{aligned} \tag{10}$$

In the posterior factorization, the first factor is a function of both θ and γ, and hides an average over $z_{1:T}$. In App. B.1 and B.2 we consider the two approximations

$$\begin{aligned} p_{\theta,\gamma}(a_{1:T}|x_{1:T})\, p_\gamma(z_{1:T}|a_{1:T}) &\approx q^{\text{D}}_\phi(a_{1:T}|x_{1:T})\, p_\gamma(z_{1:T}|a_{1:T}) \\ &= \Big\{ \prod_t q^{\text{D}}_\phi(a_t|x_t) \Big\} p_\gamma(z_{1:T}|a_{1:T})\,, \end{aligned} \tag{11}$$

2. In truth, a forward filtering backward sampling pass is required to sample $\widehat{z}_{1:T} \sim q^{\text{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T})$ jointly. These should then be used to sample $\widehat{a}_{1:T} \sim q^{\text{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T}, \widehat{z}_{1:T})$, as node a_t includes a local potential $q^*_\phi(a_t|x_t)$.

and

$$p_{\theta,\gamma}(a_{1:T}|x_{1:T})\, p_\gamma(z_{1:T}|a_{1:T}) \approx q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T})\, p_\gamma(z_{1:T}|a_{1:T})\,. \tag{12}$$

The difference between these approximations is subtle but important: the first factor in (11) factorizes, whilst the first factor in (12) incorporates an average over $z_{1:T}$ – just like (10) – and doesn't factorize. This simple observation summarizes why, in the experimental results, $\mathcal{L}^{\mathrm{U}}(\theta,\gamma,\phi)$ from (6) yields a much tighter bound than $\mathcal{L}^{\mathrm{D}}(\theta,\gamma,\phi)$ from (4).

B.1. Directed Graph Variational Approximation: Derivations

B.1.1. Evidence Lower Bound

Using $q^{\mathrm{D}}_\phi(a_{1:T}|x_{1:T}) = \prod_{t=1}^T q^{\mathrm{D}}_\phi(a_t|x_t)$ from (3), we bound the log marginal likelihood with

$$\begin{aligned}
\log p_{\theta,\gamma}(x_{1:T}) &= \int_{a_{1:T}} p_\theta(x_{1:T}|a_{1:T})\, p_\gamma(a_{1:T}) \\
&\geq \int_{a_{1:T}} q^{\mathrm{D}}_\phi(a_{1:T}|x_{1:T}) \log \frac{p_\theta(x_{1:T}|a_{1:T})\, p_\gamma(a_{1:T})}{q^{\mathrm{D}}_\phi(a_{1:T}|x_{1:T})} \\
&= \mathbb{E}_{q^{\mathrm{D}}(a_{1:T}|x_{1:T})}[\log p_\theta(x_{1:T}|a_{1:T})] - \mathrm{KL}\left(q^{\mathrm{D}}_\phi(a_{1:T}|x_{1:T}) \,\middle\|\, p_\gamma(a_{1:T})\right) \\
&= \mathcal{L}^{\mathrm{D}}(\theta,\gamma,\phi)
\end{aligned} \tag{13}$$

to yield the ELBO in (4).

B.2. Undirected Graph Variational Approximation: Derivations

To obtain the undirected model of Sec. 4, we replace the $p_\theta(x_t|a_t)$ factors in the generative model with Gaussian approximations $q^*_\phi(a_t|x_t) = \mathcal{N}(a_t; \mu^*_\phi(x_t), \Sigma^*_\phi(x_t))$, where $\mu^*_\phi(x_t)$ and $\Sigma^*_\phi(x_t)$ are neural networks returning the mean and a diagonal covariance matrix.

Before deriving the ELBO, we first present an expression for $Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})$ in (3), and describe how it is computed analytically:

$$\begin{aligned}
Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T}) &= \int_{z_{1:T}} p_\gamma(z_1) \prod_{t=2}^T p_\gamma(z_t|z_{t-1}) \int_{a_{1:T}} \prod_{t=1}^T p_\gamma(a_t|z_t)\, q^*_\phi(a_t|x_t) \\
&= \int_{z_{1:T}} p_\gamma(z_1) \prod_{t=2}^T p_\gamma(z_t|z_{t-1}) \prod_{t=1}^T \left(\int_{a_t} \mathcal{N}(a_t; Bz_t, \Sigma_a)\, \mathcal{N}(a_t; \mu^*_\phi(x_t), \Sigma^*_\phi(x_t)) \right)
\end{aligned} \tag{14}$$

$$= \int_{z_{1:T}} p_\gamma(z_1) \prod_{t=2}^T p_\gamma(z_t|z_{t-1}) \prod_{t=1}^T \mathcal{N}(0; Bz - \mu^*_\phi(x_t), \Sigma_a + \Sigma^*_\phi(x_t))\,. \tag{15}$$

The step from (14) to (15) is done by noting that the integral over a_t is the convolution of two Gaussian distributions, the density of the difference of random variables, evaluated at zero. By rewriting $\mathcal{N}(0; Bz - \mu^*_\phi(x_t), \Sigma_a + \Sigma^*_\phi(x_t)) = \mathcal{N}(\mu^*_\phi(x_t); Bz, \Sigma_a + \Sigma^*_\phi(x_t))$, (15) is the directed graph probability distribution for an LGSSM where the $\mu^*_\phi(x_1), .., \mu^*_\phi(x_T)$ are treated as point observations with unique emission noise covariances for each time step $\Sigma_a + \Sigma^*_\phi(x_t)$. The $z_{1:T}$ variables may be integrated out by one forward pass of Kalman filtering to give $Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})$.

B.2.1. Evidence Lower Bound

We lower bound $\log p_{\theta,\gamma}(x_{1:T})$ using the marginal $q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T})$, as it is defined in (5). One can also derive a lower bound using the marginal $q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T})$; we present this bound in App. B.2.2.

We first integrate out $a_{1:T}$ from the inference model in (5):

$$\begin{aligned} q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T}) &= \int_{a_{1:T}} q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}, z_{1:T}|x_{1:T}) \\ &= \frac{1}{Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})}\, p_\gamma(z_{1:T}) \prod_{t=1}^{T} \int_{a_t} q^*_\phi(a_t|x_t)\, p_\gamma(a_t|z_t) \\ &= \frac{1}{Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})}\, p_\gamma(z_{1:T}) \prod_{t=1}^{T} \mathcal{N}(\mu^*_\phi(x_t); Bz_t, \Sigma_a + \Sigma^*_\phi(x_t))\,. \end{aligned} \tag{16}$$

The evidence lower bound $\mathcal{L}^{\mathrm{U}}(\theta,\gamma,\phi)$ is given by:

$$\begin{aligned} \log p_{\theta,\gamma}(x_{1:T}) &= \log \int_{z_{1:T}} p_{\theta,\gamma}(x_{1:T}|z_{1:T}) p_\gamma(z_{1:T}) \\ &\geq \int_{z_{1:T}} q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T}) \log \frac{p_{\theta,\gamma}(x_{1:T}|z_{1:T}) p_\gamma(z_{1:T})}{q^{\mathrm{U}}_{\phi,\theta}(z_{1:T}|x_{1:T})} \\ &= \int_{z_{1:T}} q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T}) \log \frac{Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})\, p_{\theta,\gamma}(x_{1:T}|z_{1:T}) p_\gamma(z_{1:T})}{\prod_{t=1}^{T} \mathcal{N}(\mu^*_\phi(x_t); Bz_t, \Sigma_a + \Sigma^*_\phi(x_t))\, p_\gamma(z_{1:T})} \\ &= \mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}} \left[\sum_t \log p_{\theta,\gamma}(x_t|z_t) \right] \\ &\quad \underbrace{-\mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}} \left[\sum_t \log \mathcal{N}(\mu^*_\phi(x_t); Bz_t, \Sigma_a + \Sigma^*_\phi(x_t)) \right] + \log Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})}_{-\mathrm{KL}(q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T}) \,\|\, p_\gamma(z_{1:T}))} \\ &= \mathcal{L}^{\mathrm{U}}(\theta,\gamma,\phi)\,. \end{aligned} \tag{17}$$

The expectations in (17) are over $q^{\mathrm{U}}_{\phi,\gamma}(z_t|x_{1:T})$, which are found by integrating out $a_{1:T}$ and Kalman filtering and Rauch-Tung-Striebel smoothing over $z_{1:T}$. Let the output of this deterministic computation by represented with the shorthand

$$q^{\mathrm{U}}_{\phi,\gamma}(z_t|x_{1:T}) = \mathcal{N}\Big(z_t\,;\, m^t_{\phi,\gamma}(x_{1:T}),\, V^t_{\phi,\gamma}(x_{1:T})\Big) = \mathcal{N}(z_t\,;\, m_t,\, V_t)\,, \tag{18}$$

where $m^t_{\phi,\gamma}(x_{1:T})$ denotes the output of the forward-backward computation for time step t that produces the Gaussian mean of z_t, and $V^t_{\phi,\gamma}(x_{1:T})$ denotes the same output that yields the covariance of z_t. Both take all observations $x_{1:T}$ as input, and are functions of ϕ and γ. We use m_t and V_t below for uncluttered notation.

Armed with this shorthand, the first term in (17) may be written in a form that is amenable to the "reparameterization trick":

$$\mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T})} \left[\sum_t \log p_{\theta,\gamma}(x_t|z_t) \right] = \sum_t \mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}(z_t|x_{1:T})} \Big[\log p_{\theta,\gamma}(x_t|z_t) \Big]$$

$$= \sum_t \mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}(z_t|x_{1:T})} \left[\log \int_{a_t} p_\theta(x_t|a_t)\, p_\gamma(a_t|z_t) \right]$$
$$\approx \sum_t \frac{1}{N_z} \sum_{n_z=1}^{N_z} \log \left(\frac{1}{N_a} \sum_{n_a=1}^{N_a} p_\theta(x_t|a_t^{(n_z,n_a)}) \right) \tag{19}$$

where we set $N_z = N_a = 1$ and the reparameterization trick is used for the samples in (19):

$$z_t^{(n_z)} \sim q^{\mathrm{U}}_{\phi,\gamma}(z_t|x_{1:T}), \qquad z_t^{(n_z)} = m_t + \mathrm{chol}(V_t)\epsilon^{(n_z)}, \tag{20}$$
$$a_t^{(n_z,n_a)} \sim p_\gamma(a_t|z_t^{(n_z)}), \qquad a_t^{(n_z,n_a)} = B z_t^{(n_z)} + \mathrm{chol}(\Sigma_a)\epsilon^{(n_z,n_a)}\ . \tag{21}$$

All $\epsilon^{(\cdot)}$ values are independent $\mathcal{N}(\epsilon^{(\cdot)}; 0, I)$ samples.

Returning to (17), the last $\log Z^{\mathrm{U}}_{\phi,\theta}(x_{1:T})$ term is found analytically, as described in (15) and the discussion around it; and the middle term is the cross entropy of Gaussians which can be computed in closed form.

Finally, as an aside, we show below that the KL divergence between $q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T})$ and $p_\gamma(z_{1:T})$ corresponds to that annotated in (17):

$$\begin{aligned}
&\mathrm{KL}\Big(q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T}) \,\Big\|\, p_\gamma(z_{1:T}) \Big) \\
&= \int_{z_{1:T}} q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T}) \log \frac{q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T})}{p_\gamma(z_{1:T})} \\
&= \int_{z_{1:T}} q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T}) \log \frac{\prod_{t=1}^T \mathcal{N}(\mu^*_\phi(x_t); Bz_t, \Sigma_a + \Sigma^*_\phi(x_t))\, p_\gamma(z_{1:T})}{Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})\, p_\gamma(z_{1:T})} \\
&= \mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}} \left[\sum_t \log \mathcal{N}(\mu^*_\phi(x_t); Bz_t, \Sigma_a + \Sigma^*_\phi(x_t)) \right] - \log Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})\,.
\end{aligned} \tag{22}$$

Notice that $p_\theta(z_{1:T})$ cancels out in the second last line of (22).

B.2.2. Evidence Lower Bound II

One might also consider an alternative ELBO than $\mathcal{L}^{\mathrm{U}}(\theta, \gamma, \phi)$ from (6), by using the marginal $q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T})$ to derive the ELBO in (24). (When $q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}, z_{1:T}|x_{1:T})$ is used to construct an ELBO, the same bound as (24) is obtained.) We first marginalize (5) over $z_{1:T}$:

$$\begin{aligned}
q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T}) &= \frac{1}{Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})} \left\{ \prod_{t=1}^T q^*_\phi(a_t|x_t) \right\} \int_{z_{1:T}} p_\gamma(a_{1:T}|z_{1:T})\, p_\gamma(z_{1:T}) \\
&= \frac{1}{Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})} \left\{ \prod_{t=1}^T q^*_\phi(a_t|x_t) \right\} p_\gamma(a_{1:T})\,.
\end{aligned} \tag{23}$$

Another lower bound than the one we encountered in App. B.2.1 is obtained with

$$\log p_{\theta,\gamma}(x_{1:T}) = \log \int_{a_{1:T}} p_\theta(x_{1:T}|a_{1:T})\, p_\gamma(a_{1:T})$$

$$
\begin{aligned}
&\geq \int_{a_{1:T}} q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T}) \log \frac{p_\theta(x_{1:T}|a_{1:T}) p_\gamma(a_{1:T})}{q^{\mathrm{U}}_{\phi,\theta}(a_{1:T}|x_{1:T})} \\
&= \int_{a_{1:T}} q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T}) \log \frac{Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})\, p_\theta(x_{1:T}|a_{1:T}) p_\gamma(a_{1:T})}{\prod_{t=1}^T q^*_\phi(a_t|x_t)\, p_\gamma(a_{1:T})} \\
&= \mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}}\left[\sum_t \log p_\theta(x_t|a_t)\right] \underbrace{- \mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}}\left[\sum_t \log q^*_\phi(a_t|x_t)\right]}_{-\mathrm{KL}(q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T}) \,\|\, p_\gamma(a_{1:T}))} + \log Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T}) \\
&= \mathcal{L}^{\mathrm{U}_2}(\theta,\gamma,\phi)\,. \qquad (24)
\end{aligned}
$$

One can determine the marginals $q^{\mathrm{U}}_{\phi,\gamma}(a_t|x_{1:T})$ with a single forward-backward message passing procedure, as the graph is a tree. The first term of the three terms in (24) may be evaluated using Monte-Carlo integration (employing the "reparameterization trick") and the final two terms are computed analytically. This ELBO has a pleasing interpretation: maximizing it maximizes the (approximate) marginal likelihood $\log Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})$ whilst simultaneously minimizing the average *local* encoding-decoding cost $\sum_t \mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}}[\log p_\theta(x_t|a_t) - \log q^*(a_t|x_t)]$.

As an aside, again, we show below that the KL divergence between $q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T})$ and $p_\gamma(a_{1:T})$ corresponds to that annotated in (24):

$$
\begin{aligned}
&\mathrm{KL}\Big(q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T}) \,\Big\|\, p_\gamma(a_{1:T})\Big) \\
&= \int_{a_{1:T}} q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T}) \log \frac{q^{\mathrm{U}}_{\phi,\gamma}(a_{1:T}|x_{1:T})}{p_\gamma(a_{1:T})} \\
&= \int_{a_{1:T}} q^{\mathrm{U}}_{\phi,\gamma}(z_{1:T}|x_{1:T}) \log \frac{\prod_{t=1}^T q^*_\phi(a_t|x_t)\, p_\gamma(a_{1:T})}{Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})\, p_\gamma(a_{1:T})} \\
&= \mathbb{E}_{q^{\mathrm{U}}_{\phi,\gamma}}\left[\sum_t \log q^*_\phi(a_t|x_t)\right] - \log Z^{\mathrm{U}}_{\phi,\gamma}(x_{1:T})\,. \qquad (25)
\end{aligned}
$$

Notice that $p_\theta(a_{1:T})$ cancels out in the second last line of (25).

Appendix C. Further Experimental Results

C.1. Inferred Latent Positions

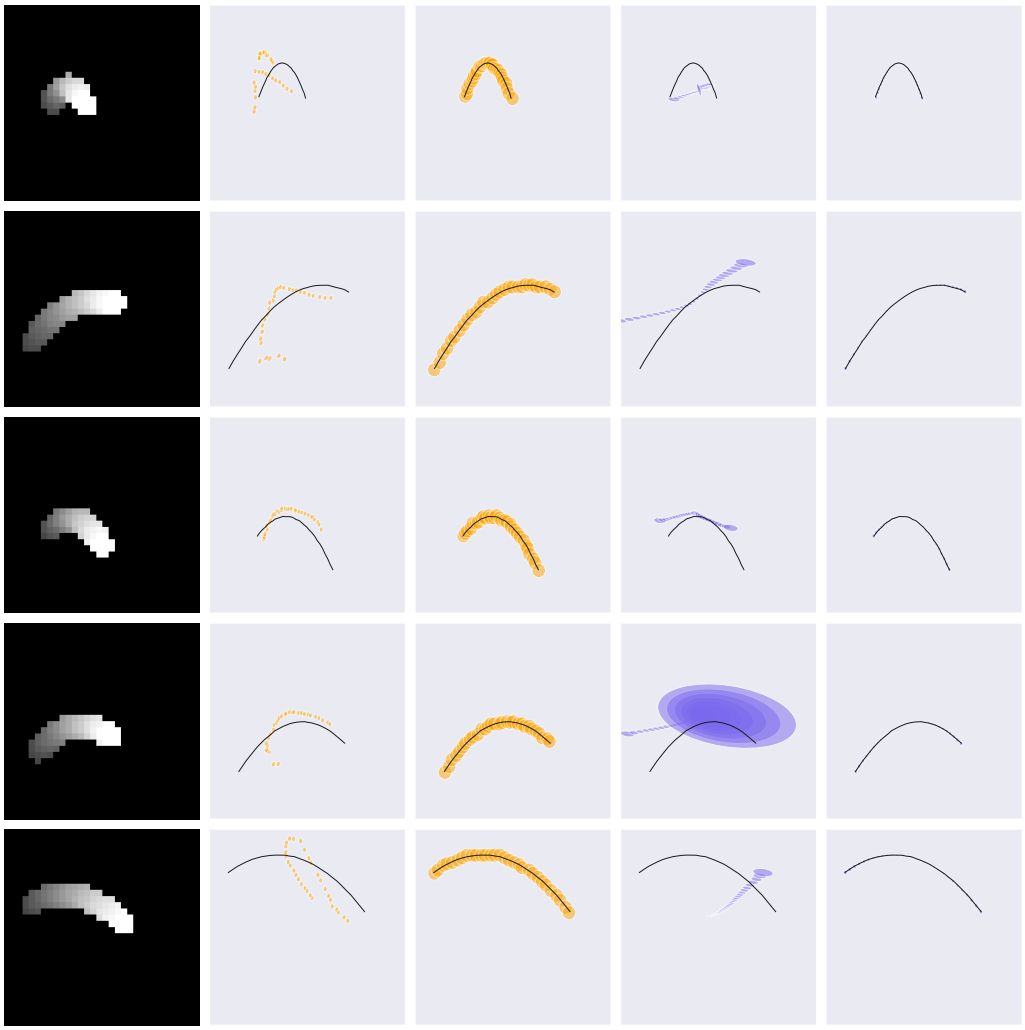

Figure 4: *Column 1*: Observed video sequence. *Columns 2, 3*: Directed model inferred positions with $\beta_0 = 1$ (standard training) *(column 2)* and annealed from $\beta_0 = 100$ *(column 3)*. The shaded regions represent two standard deviations. *Columns 4, 5*: Undirected model inferred positions with $\beta_0 = 1$ *(column 4)* and $\beta_0 = 100$ *(column 5)*. The shaded regions represent 20 standard deviations, posterior approximation variance is much much smaller than for the directed model. We plot the rotated inferred position using the minimum squared error linear transformation from (8). For both models, standard training leads to learning a latent domain that does not follow the imposed Newtonian parabolic motion. Annealing from $\beta_0 = 100$ encourages the model to learn a mapping to the latent domain that corresponds to reality.

C.2. Varying β Annealing and LGSSM Parameters

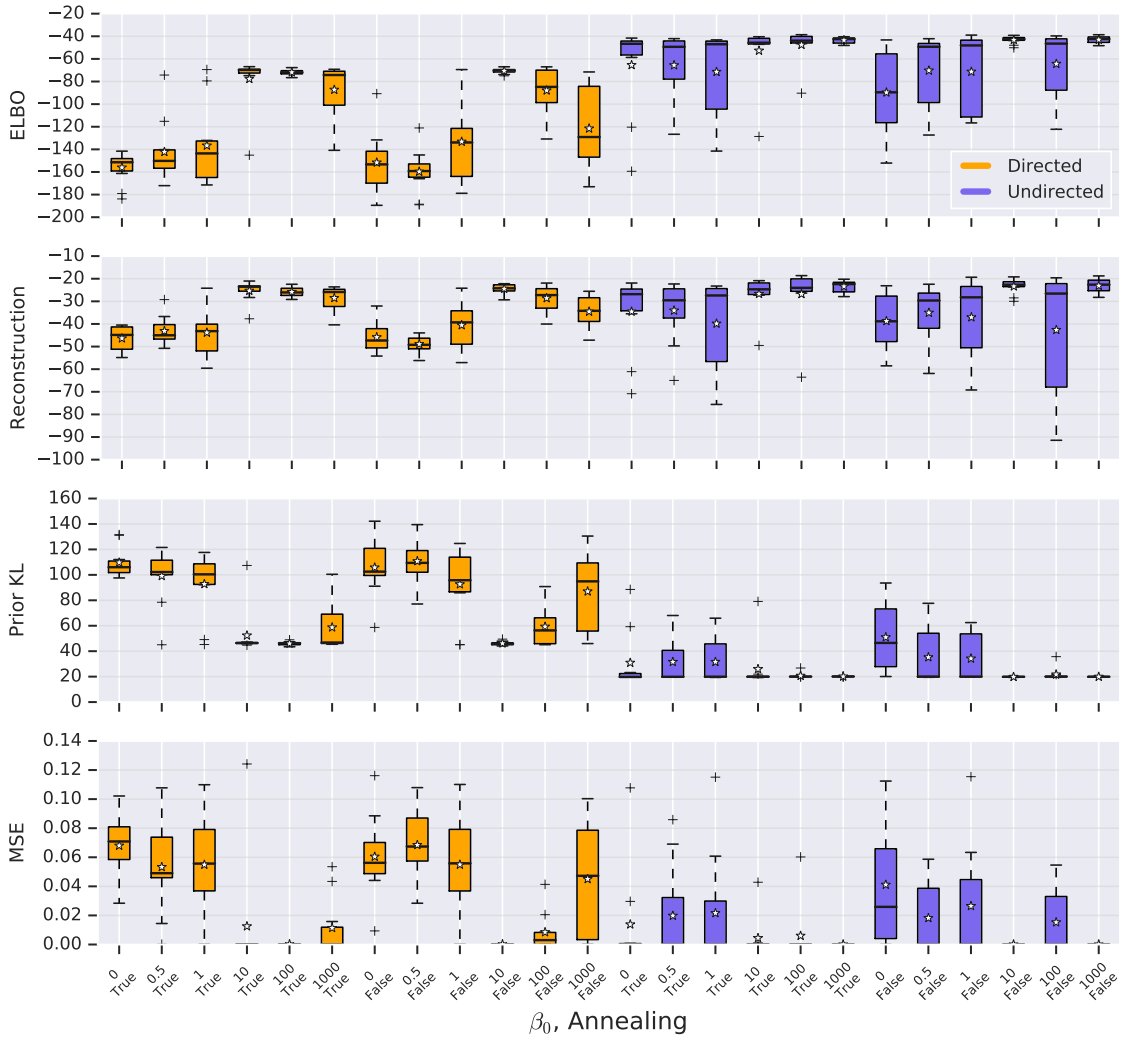

Figure 5: Each inference model was trained with and without annealing the β_0 as described in App. A.2. For all plots, within results for each model, left hand error bars are after training with annealing (True), right hand are without annealing (False). *Plot 1.* The directed model with annealing achieves the best lower bound when $\beta = 10, 100$, without annealing, only $\beta_0 = 10$ is best suggesting annealing reduces sensitivity on the training hyper parameter β_0. For the undirected model, with the exception of $\beta_0 = 100$ where 3 of the 10 seeds resulted in an unstable decoder, any value of $\beta_0 \geq 10$ yields the same bound regardless of annealing. *Plot 2.* As above the reconstruction error for the directed model depends on annealing and does not for the undirected model. *Plots 3 and 4.* The Prior KL divergence and ground truth prediction are uniformly lower for the undirected model for all training schedules.

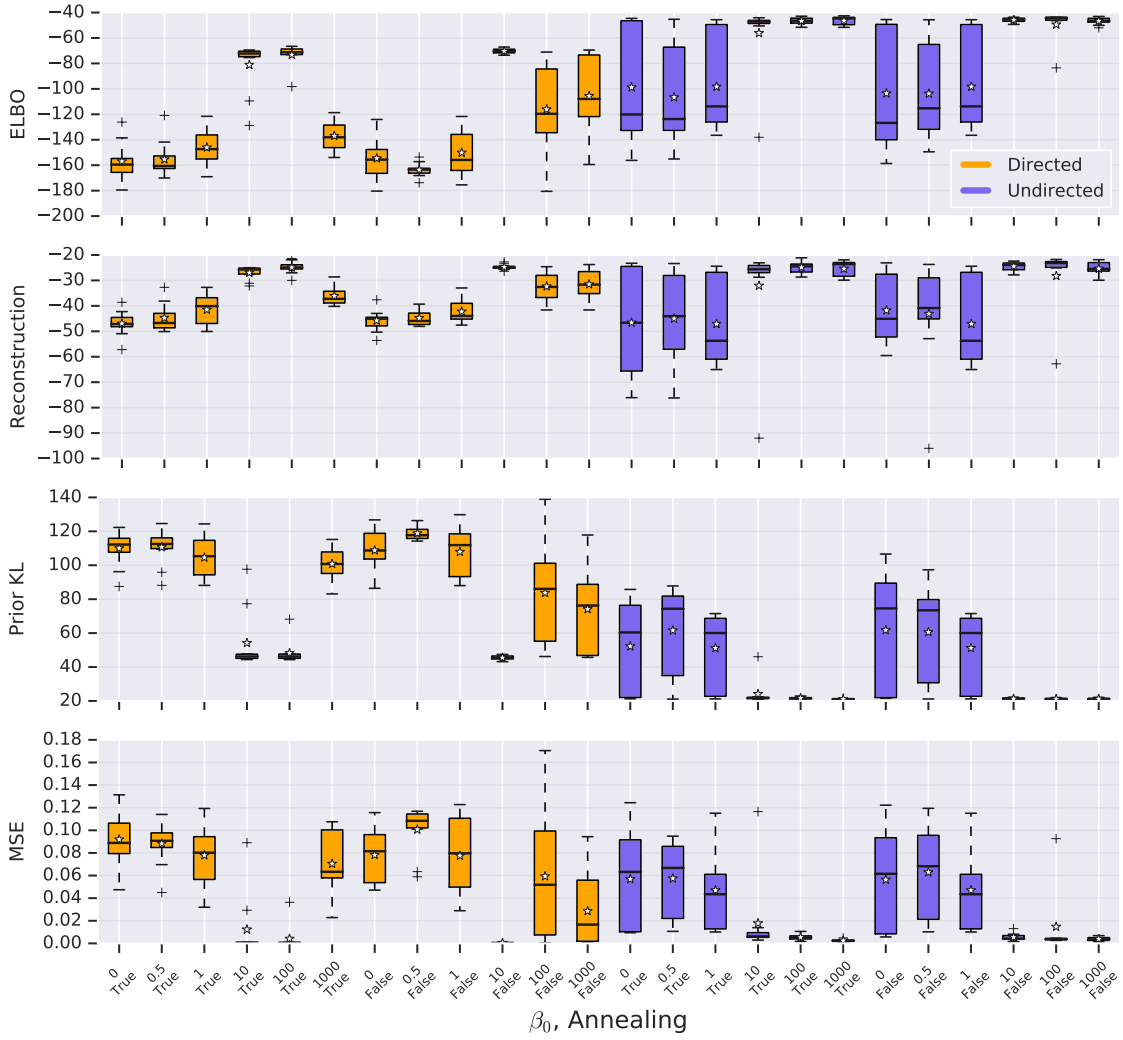

Figure 6: To evaluate the sensitivity of each inference model with respect to the LGSSM transition and emission matrices, we repeated all the experimented learning both matrices. *Plot 1.* As with fixed LGSSM parameters, annealing reduces the sensitivity of the directed model upon β_0 and the undirected model achieves a greater lower bound for all training schedules and achieves best performance with $\beta_0 \geq 10$ regardless of annealing. *Plot 2.* For all $\beta_0 \leq 1$, the directed model learns a more stable decoder. *Plot 3.* Comparing with Plot 2, we see that the improvement in the lower bound for the undirected model comes from matching the learned prior dynamics. *Plot 4.* For all values of β_0, the undirected model never learns to completely recover the ground truth. For the undirected model, fixed dynamics and $\beta_0 \geq 10$ is required to learn a physical latent domain whereas for the directed model, a narrower range of β_0 and annealing is required but not fixed LGSSM parameters.

C.3. Variational Posteriors and Prior Divergence per Frame

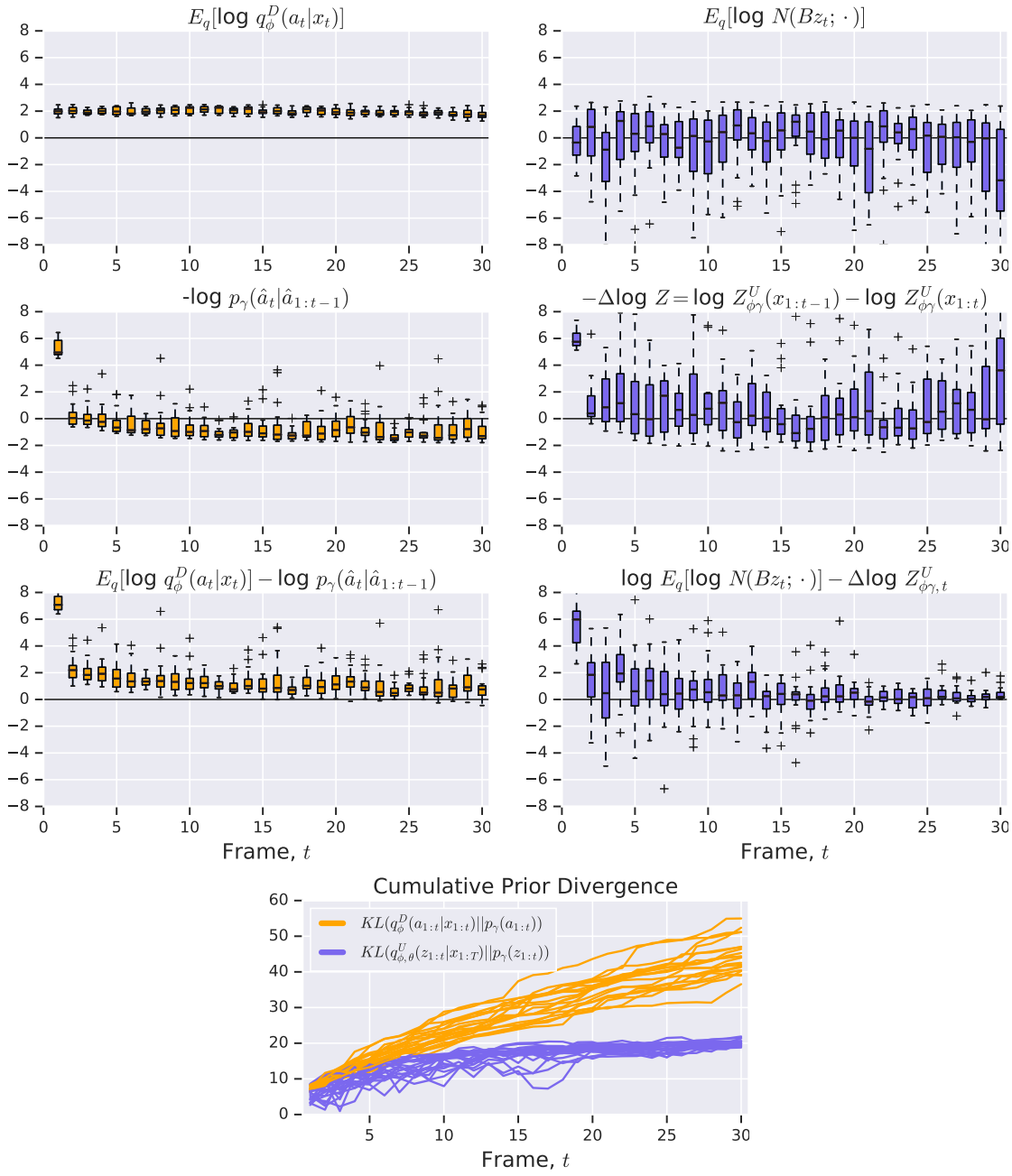

Figure 7: Both models were trained with annealing from $\beta_0 = 100$. *Top left.* The directed model posterior entropy per frame. *Middle left.* Per step likelihood from filtering. *Bottom left.* The sum of above two plots is consistently positive leading to increasing divergence over time. *Top right.* The per-frame cross-entropy term of the ELBO for the directed model. *Middle right.* The normalizing constant per frame. *Bottom right.* Summing the above plots leads to a series that shrinks over time. *Bottom.* The cumulative sum for each video in the batch. The directed model divergence consistently increases with video length unlike the undirected model that exploits dynamics.