Master Thesis

Human activity recognition and prediction from motion capture and additional sensors

Maxime Chaveroche

Academic Year 2016–2017

Final year internship done in partnership with

INRIA

in preparation for the engineering diploma of TELECOM Nancy

Internship supervisor: Serena Ivaldi, Francis Colas

Academic supervisor: Bruno Pinçon

1 State of the art

1.1 Detailed background of our method

1.1.1 VAE

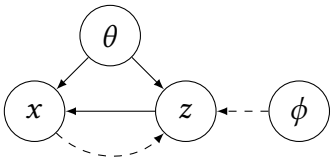


Figure 1.1: VAE representation consisting of the superposition of two Bayesian models. Solid lines denote the generative model with parameters θ, dashed lines denote the inference model with parameters ϕ.

A VAE [5] is an auto-encoder based on *Variational Inference* (VI) [1] that attempts to find a reduced representation of data generalizable to variants of that data.

Let $[x_k]_{k=1}^K$ be a dataset of K i.i.d. samples of some continuous observation variable x of unknown distribution. We assume that the data are generated by some random process, involving a latent continuous random variable z and parametric families of distributions $p_\theta(x|z)$ and $p_\theta(z)$:

$$x \sim \int_z p_{\theta^*}(x|z)p_{\theta^*}(z)dz.$$

where θ^* is the set of parameters of the parametric distribution $p_{\theta^*}(x)$. In our case, θ will be represented by weights and biases of a *decoder* neural network. z is a latent space that can be chosen arbitrarily; in VAE, the distributions are chosen to be Gaussian:

$$p_\theta(z) = \mathcal{N}(0, I), \text{ and } p_\theta(x|z) = \mathcal{N}(\mu_x, \sigma_x^2 I),$$

However, we don't know the distribution of $p_\theta(x)$, and so the distribution of $p_\theta(z|x)$ as $p_\theta(z|x) = \frac{p_\theta(z,x)}{p_\theta(x)}$. Therefore, [5] introduced a recognition model $q_\phi(z|x)$ designed to approximate the intractable true posterior $p_\theta(z|x)$. We could then constrain $q_\phi(z|x)$ to the same form as $p_\theta(x|z)$:

$$q_\phi(z|x) = \mathcal{N}(\mu_z, \sigma_z^2 I), \tag{1.1}$$

where ϕ is the set of parameters of the parametric distribution $q_\phi(z|x)$. In our case, ϕ will be represented by weights and biases of an *encoder* neural network.

This model, illustrated by figure 1.1, can then simultaneously encode and decode x. Yet, we have to learn θ and ϕ in a way that maximizes the likelihood of $[x_k]_{k=1}^K$.

Training in VAE aims both at recovering parameters θ as close as possible to the ideal parameters θ^* yielding the proper distribution over x, and finding parameters ϕ making $q_\phi(z|x)$ as close as possible to the intractable $p_\theta(z|x)$. This can be achieved by minimizing the Kullback-Leibler divergence $D_{KL}(q_\phi(z|x) \parallel p_\theta(z|x))$ between these distributions. The standard approach of variational inference is to notice that we could decompose $\ln(p_\theta([x_k]_{k=1}^K))$, isolate $D_{KL}(q_\phi(z|x_k) \parallel p_\theta(z|x_k))$ and minimize it by maximizing a lower bound on $\ln(p_\theta(x_k))$:

$$\begin{aligned} \ln(p_\theta([x_k]_{k=1}^K)) &= \ln(\prod_{k=1}^K p_\theta(x_k)) \\ &= \sum_{k=1}^K ln(p_\theta(x_k)) \end{aligned} \tag{1.2}$$

$$\begin{aligned} \ln(p_\theta(x_k)) &= \int_z q_\phi(z|x_k) \ln(p_\theta(x_k)) dz \\ &= \int_z q_\phi(z|x_k) \ln(\frac{p_\theta(z,x_k)}{q_\phi(z|x_k)}) + q_\phi(z|x_k) \ln(\frac{q_\phi(z|x_k)}{p_\theta(z|x_k)})\ dz \\ &= ELBO + D_{KL}(q_\phi(z|x_k) \parallel p_\theta(z|x_k)) \end{aligned} \tag{1.3}$$

where *ELBO* is the Evidence Lower BOund.

As the actual likelihood on the observation $p_{\theta^*}(x_k)$ is a constant and a KL-divergence is always non-negative, minimizing that KL-divergence can be done by maximizing the *ELBO*. In terms of training the neural network, it means using $-ELBO$ as the loss function to be minimized. This bound can in turn be split into *reconstruction* (or decoder) error and *generalization* (or encoder) error:

$$\begin{aligned} -ELBO &= -\int_z q_\phi(z|x_k) \ln(\frac{p_\theta(z,x_k)}{q_\phi(z|x_k)})\ dz \\ &= -\int_z q_\phi(z|x_k) \ln(p_\theta(x_k|z)) + q_\phi(z|x_k) \ln(\frac{p_\theta(z)}{q_\phi(z|x_k)})\ dz \\ &= -\mathbb{E}_{q_\phi(z|x_k)}[\ln(p_\theta(x_k|z))] + D_{KL}(q_\phi(z|x_k) \parallel p_\theta(z)) \end{aligned} \tag{1.4}$$

The error on the left hand side requires precision in reconstruction from the network by making it maximize the average likelihood $p_\theta(x_k|z)$ on z inferred by the recognition model $q_\phi(z|x_k)$. The error on the right hand side requires generalization from the network and balances the first error which could result in overfitting otherwise.

With our Gaussian assumptions, we can show that:

$$D_{KL}(q_\phi(z|x_k) \parallel p_\theta(z)) = \frac{1}{2} \sum_{i=1}^{d_z} (\sigma_{z,k,i}^2 + \mu_{z,k,i} - \ln(\sigma_{z,k,i}^2) - 1),$$

where d_z is the number of dimensions of z, and:

$$-\mathbb{E}_{q_\phi(z|x_k)}[\ln(p_\theta(x_k|z))] \approx \frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{d_x} \frac{1}{2} \left[\ln(2\pi) + \ln(\sigma_{x,k,l,i}^2) + \frac{(x_{k,i} - \mu_{x,k,l,i})^2}{\sigma_{x,k,l,i}^2} \right] \tag{1.5}$$

where d_x is the number of dimensions of x, and the expectation is approximated by sampling L values:

$$\forall l \in [\![1, L]\!], \quad z_{k,l} = \mu_{z,k} + \sigma_{z,k} \odot \epsilon_l \tag{1.6}$$

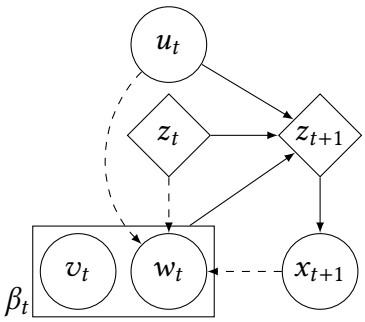


Figure 1.2: DVBF representation consisting of the superposition of two Bayesian models. Dependencies to θ and ϕ have been omitted for the sake of clarity. Solid lines denote the generative model with parameters θ, dashed lines denote the inference model with parameters ϕ. Diamond nodes indicate a deterministic dependency on parent nodes. A rectangle designates the joint distribution of the variables inside, here β_t.

for $z_{k,l} \sim$ (1.1) from which derive $\mu_{x,k,l}$ and $\sigma_{x,k,l}$, where $\epsilon \sim \mathcal{N}(0, I)$ and $\odot$ is the element-wise product. (1.6) is known as the *reparametrization trick* and allows the back-propagation algorithm to reach the encoder weights.

The influence of σ_z on (1.5) shows the difference of VAE with respect to more standard auto-encoders, as this term allow for the learning of a locally symmetrical and continuous latent space, and thus the importance of the *generalization* error for feature extraction.

1.1.2 DVBF

Let us now consider a set of data sequences $[[x_{k,t}]_{t=1}^{T}]_{k=1}^{K}$. We will then consider a single sample k and use only the notation $\cdot_t$ as we are mostly interested in time here.

DVBF [4] was introduced to handle time-dependency in data sequences. Latent variables are supposed to follow a dynamics equation expressed as:

$$z_{t+1} = g(z_t, u_t, \beta_t). \tag{1.7}$$

where g is a deterministic transition function, u_t is a command at time t and β_t is a set of parameters written as $\beta_t = (w_t, v_t)$, where v_t are fixed universal transition parameters, and w_t is a stochastic variable representing a sample-specific process noise which can be inferred from x_t.

In [4], the authors also assumed that all information about the current observation x_t is contained in the current latent variable z_t. Coupled with the Markov assumption, we can write:

$$\begin{cases} p_\theta(x_{1:T}|z_{1:T}, u_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t|z_t), \\ p_\theta(z_{1:T}|\beta_{1:T}, u_{1:T}) = p_\theta(z_1) \prod_{t=2}^{T} p_\theta(z_t|z_{t-1}, u_{t-1}, \beta_{t-1}). \end{cases} \tag{1.8}$$

where $\cdot_{1:T} = (\cdot_1, \cdot_2, ..., \cdot_T)$ and T is the data sequence length.

The representation in figure 1.2 describes the Bayesian models used here. Under these conditions,

they calculated a variational lower bound specific to DVBF :

$$\begin{aligned} p_\theta(x_{1:T}|u_{1:T},z_1) &= \int_{z_{2:T}} p_\theta(x_{1:T}|z_{1:T},u_{1:T})p_\theta(z_{2:T}|u_{1:T})dz_{2:T} \\ &= \int_{z_{2:T}}\int_{\beta_{1:T}} p_\theta(x_{1:T}|z_{1:T},u_{1:T})p_\theta(z_{2:T}|\beta_{1:T},u_{1:T})p_\theta(\beta_{1:T})d\beta_{1:T}dz_{2:T} \\ &= \int_{z_{2:T}}\int_{\beta_{1:T}} p_\theta(\beta_{1:T})\prod_{t=1}^{T} p_\theta(x_t|z_t)\prod_{t=2}^{T} p_\theta(z_t|z_{t-1},u_{t-1},\beta_{t-1})d\beta_{1:T}dz_{2:T} \end{aligned} \tag{1.9}$$

Given the deterministic transition (1.7), they noticed that $\prod_{t=2}^{T} p_\theta(z_t|z_{t-1},u_{t-1},\beta_{t-1})$ was a product of Dirac distributions and thus:

$$p_\theta(x_{1:T}|u_{1:T},z_1) = \int_{\beta_{1:T}} p_\theta(\beta_{1:T})p_\theta(x_t|z_1)\prod_{t=2}^{T} p_\theta(x_t|\hat{z}_t)\ d\beta_{1:T} \tag{1.10}$$

where $\hat{z}_t = g(z_{t-1},u_{t-1},\beta_{t-1})$. Then, taking the logarithm of $p_\theta(x_{1:T}|u_{1:T},z_1)$:

$$\begin{aligned} &\ln(p_\theta(x_{1:T}|u_{1:T},z_1)) \\ &= \ln\left(\int_{\beta_{1:T}} \frac{q_\phi(\beta_{1:T}|x_{1:T},u_{1:T},z_1,\hat{z}_{2:T})}{q_\phi(\beta_{1:T}|x_{1:T},u_{1:T},z_1,\hat{z}_{2:T})}\ p_\theta(\beta_{1:T})p_\theta(x_t|z_1)\prod_{t=2}^{T} p_\theta(x_t|\hat{z}_t)\ d\beta_{1:T}\right) \\ &\geq \int_{\beta_{1:T}} q_\phi(\beta_{1:T}|x_{1:T},u_{1:T},z_1,\hat{z}_{2:T})\ln\left(\frac{p_\theta(\beta_{1:T})}{q_\phi(\beta_{1:T}|x_{1:T},u_{1:T},z_1,\hat{z}_{2:T})}p_\theta(x_t|z_1)\prod_{t=2}^{T} p_\theta(x_t|\hat{z}_t)\right) d\beta_{1:T} = LB \end{aligned} \tag{1.11}$$

where:

$$LB = \mathbb{E}_{q_\phi(\beta_{1:T}|x_{1:T},u_{1:T},z_1,\hat{z}_{2:T})}\left[\ln\left(p_\theta(x_t|z_1)\prod_{t=2}^{T} p_\theta(x_t|\hat{z}_t)\right)\right] - D_{KL}(q_\phi(\beta_{1:T}|x_{1:T},u_{1:T},z_1,\hat{z}_{2:T})\ \|\ p_\theta(\beta_{1:T})) \tag{1.12}$$

As you can see, that lower bound doesn't link to the minimization of $D_{KL}(q_\phi(z|x)\ \|\ p_\theta(z|x))$, but proposes a way to maximize the likelihood of $p_\theta(x_{1:T}|u_{1:T},z_1)$, embedding transition parameters learning in the encoding process of the VAE.

1.1.3 VAE-DMP

In [2], Chen et al. used DMP as a dynamics model of the latent space in DVBF. They included a system noise $\epsilon_t = \epsilon \boldsymbol{w}_{\epsilon,t}$, where $\epsilon \sim \mathcal{N}(0,\Sigma_\epsilon)$, in the DMP equation and defined their choice of finite difference approximations of first and second order derivatives:

$$\begin{cases} \ddot{z}_{t+1} = \tau\left[\alpha(\beta(z_{goal}-z_t)-\dot{z}_t)+f_t+\epsilon_t\right] \\ \ddot{z}_{t+1} = \frac{\dot{z}_{t+1}-\dot{z}_t}{dt} \\ \dot{z}_{t+1} = \frac{z_{t+1}-z_t}{dt}, \end{cases} \tag{1.13}$$

where f_t is shaped as a continuous weighted sum of Gaussian basis functions and is deterministically inferred by a MLP (detailed in [2]) that takes $x_{1:T}$ as input. This system of equations can be reshaped into the following linear form:

$$\begin{pmatrix} z_{t+1} \\ \dot{z}_{t+1} \end{pmatrix} = \begin{pmatrix} 1-\tau dt^2\alpha\beta & \tau dt^2\alpha + dt \\ -\alpha\beta\tau dt & 1-\alpha\tau dt \end{pmatrix}\begin{pmatrix} z_t \\ \dot{z}_t \end{pmatrix} + b \tag{1.14}$$

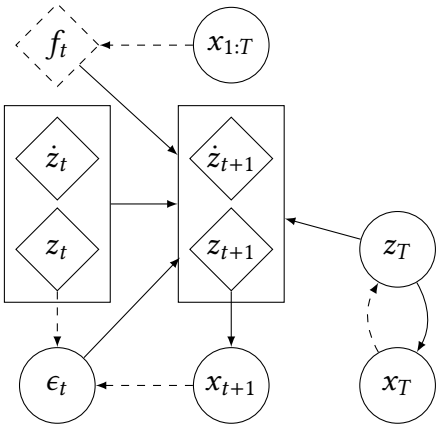


Figure 1.3: VAE-DMP representation consisting of the superposition of two Bayesian models. Dependencies to θ and ϕ have been omitted for the sake of clarity. Solid lines denote the generative model with parameters θ, dashed lines denote the inference model with parameters ϕ. Diamond nodes indicate a deterministic dependency on parent nodes. Diamond dashed node indicates a deterministic dependency on parent nodes for the inference model but not the generative one. A rectangle designates the joint distribution of the variables inside, here $(z_t, \dot{z}_t)$.

with $b = (dt, 1)^\intercal \ (\alpha\beta z_{goal} + f_t + \epsilon_t)\tau dt$. This transition model thus gives the system the notion of acceleration and forces it to be continuous through f_t.

Contrary to standard VAE, in VAE-DMP z_t is inferred by the dynamics model, along with its first-order derivative $\dot{z}_t$, at the exception of the first frame, which has no predecessor, and the last frame which is needed in (1.14). Therefore, the encoding process $q_\phi(z_1, \dot{z}_1 | x_{1:M})$, where $1 < M \leq T$, is used for the first frame, and the standard one $q_\phi(z_T | x_T)$ is used for the last frame of the architecture.

The noise is assumed inferred by a Gaussian distribution: $\epsilon_t \sim q_\phi(\epsilon_t | x_{t+1}, z_t) = \mathcal{N}(\mu_{\epsilon,t}, \Sigma_{\epsilon,t})$.

Equation (1.14) in VAE-DMP replaces equation (1.7) of DVBF: z_t is now $(z_t, \dot{z}_t)$, the command u_t of DVBF is now the joint distribution of the forcing term and the goal in latent space (f_t, z_T), and the transition parameters β_t the noise ϵ_t:

$$z_{t+1} = g(z_t, \dot{z}_t, z_T, f_t, \epsilon_t). \tag{1.15}$$

The corresponding representation can be found in figure 1.3. Noting that $f_{1:T}$ are deterministically inferred from $x_{1:T}$, the variational lower bound becomes:

$$\begin{aligned} &\ln(p_\theta(x_{1:T} | f_{1:T}, z_T, z_1, \dot{z}_1)) \\ &\geq \mathbb{E}_{q_\phi(\epsilon_{1:T} | x_{1:T}, z_T, z_1, \dot{z}_1, \hat{z}_{2:T}, \hat{\dot{z}}_{2:T})} \left[\ln(p_\theta(x_1 | z_1) p_\theta(x_{2:T} | \hat{z}_{2:T})) \right] \\ &\quad - D_{KL}(q_\phi(\epsilon_{1:T} | x_{1:T}, z_T, z_1, \dot{z}_1, \hat{z}_{2:T}, \hat{\dot{z}}_{2:T}) \parallel p_\theta(\epsilon_{1:T})) \end{aligned} \tag{1.16}$$

where $q_\phi(\epsilon_{1:T} | x_{1:T}, z_T, z_1, \dot{z}_1, \hat{z}_{2:T}, \hat{\dot{z}}_{2:T}) = q_\phi(\epsilon_{1:T} | x_{1:T}, z_1, \hat{z}_{2:T})$ given their inference model.

We note that the parameters of the prior $p_\theta(\epsilon_t)$ are not fixed values in [2], which lets the D_{KL} calculation and even purpose unclear, since its regularization power relies on fixed constraints.

2 Method

2.1 VTSFE

As said before, VTSFE is inspired by VAE-DMP. We use a similar generative model and the same assumptions on f_t but propose here a new noise inference model in subsection 2.1.1 and a new transition model in subsection 2.1.2, as well as a new lower bound for that model which can be found in subsection 2.1.3.

2.1.1 Adapted noise inference

As implied before, the inference model of DVBF has been reused in VAE-DMP. However, the variable substitution between z_t and $(z_t, \dot{z}_t)$ and the dependency on z_T of the transition model should have affected the noise inference which is supposed to fill the gap remaining between information contained in latent space and the one in observation space. Besides, knowing x_{t+1} alone isn't enough to deterministically infer f_t. Therefore, we propose the following noise inference instead:

$$\epsilon_t \sim q_\phi(\epsilon_t | f_t, x_{t+1}, z_t, \dot{z}_t, z_T), \tag{2.1}$$

if you choose to use $\dot{z}_t$, or:

$$\epsilon_t \sim q_\phi(\epsilon_t | f_t, x_{t+1}, z_t, z_{t-1}, z_T), \tag{2.2}$$

otherwise. The corresponding Bayesian model is illustrated in figure 2.1.

Moreover, let's assume that the prior $p_\theta(\epsilon_t)$ of our model is a gaussian white noise that doesn't depend on t. We consider σ_{scale} the scaling term for all noise ϵ_t, that also affects its inference mean knowing x_t. We could then make the following assumptions:

$$\begin{aligned} q_\phi(\epsilon_t | f_t, x_{t+1}, z_t, z_{t-1}, z_T) &= \mathcal{N}(\sigma_{scale} \odot \mu_{\epsilon,t},\ \sigma^2_{scale} I \sigma^2_{\epsilon,t} I) \\ p_\theta(\epsilon_t) &= \mathcal{N}(0, \sigma^2_{scale} I) \end{aligned} \tag{2.3}$$

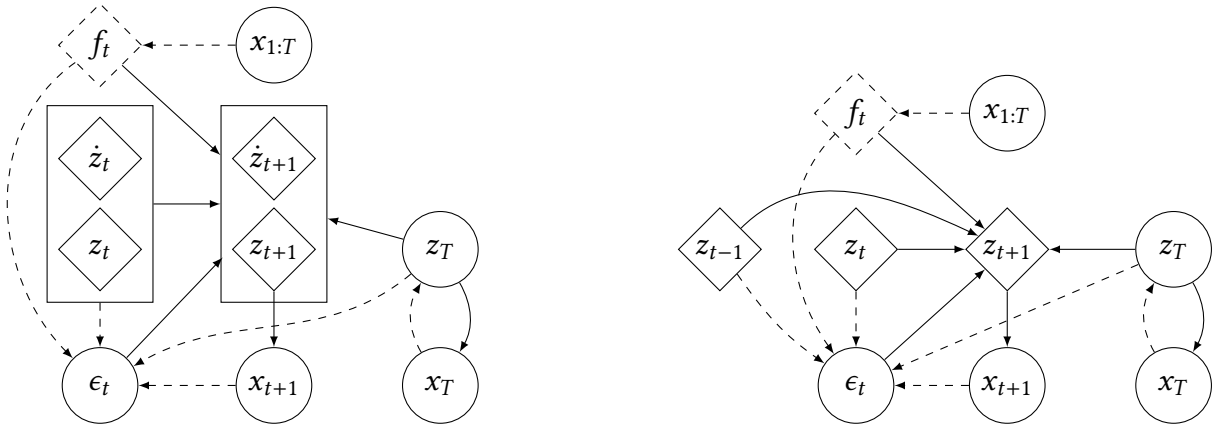
Actually, as $\mu_{\epsilon,t}$, σ_{scale} and $\sigma_{\epsilon,t}$ are all inferred by our neural network, this formulation is equivalent to:

$$\begin{aligned} q_\phi(\epsilon_t | f_t, x_{t+1}, z_t, z_{t-1}, z_T) &= \mathcal{N}(\mu_{\epsilon,t},\ \sigma^2_{\epsilon,t} I) \\ p_\theta(\epsilon_t) &= \mathcal{N}(0, \sigma^2_{scale} I) \end{aligned} \tag{2.4}$$

But (2.3) leads to a lighter Kullback-Leibler divergence expression (2.16).

2.1.2 Lighter transition model

DMP is a well-known framework used to learn trajectories, hence its use in [2] where they tried to build a good feature extractor with the VAE latent space and to create a model able to generate



(a) (b)

Figure 2.1: Our VTSFE representation consisting of the superposition of two Bayesian models (a) with $\dot{z}_t$, (b) with z_{t-1}. Dependencies to θ and ϕ have been omitted for the sake of clarity. Solid lines denote the generative model with parameters θ, dashed lines denote the inference model with parameters ϕ. Diamond nodes indicate a deterministic dependency on parent nodes. Diamond dashed node indicates a deterministic dependency on parent nodes for the inference model but not the generative one. A rectangle designates the joint distribution of the variables inside, here $(z_t, \dot{z}_t)$.

trajectories based on it. However, it could save some complexity in time and some hyperparameters if we made some adaptations to our task which is only to create a good feature extractor to be able to recognize and predict trajectories.

To do so, we only need DMP as a space-time continuity constraint, i.e. a constraint that gives to the network the notion of continuous acceleration. Therefore, we could remove the point-attractor aspect of DMP, removing the hyperparameters α, β and τ, as well as the appearance of z_T in its equation. Thus, in association with the central finite difference approximator of acceleration (which is more accurate than the backward and forward variants), (1.13) simply becomes:

$$\begin{cases} \ddot{z}_t = f_t + \epsilon_t \\ \ddot{z}_t = \frac{z_{t+1} - 2z_t + z_{t-1}}{dt^2} \end{cases} \tag{2.5}$$

And (1.14) then becomes:

$$z_{t+1} = (f_t + \epsilon_t)dt^2 + 2z_t - z_{t-1} \tag{2.6}$$

where dt doesn't even act as a hyperparameter neither since f_t and ϵ_t are learned and completely artificial.

The new transition model, formerly (1.15), becomes:

$$z_{t+1} = g(z_t, z_{t-1}, f_t, \epsilon_t), \tag{2.7}$$

and it has several benefits. First, it removes the model parametrization, which prevents us from doing an expensive grid search. Then, it greatly alleviates the complexity of our model both at transition time and at loss computation time, since it removes the need to sample the z_T prior.

2.1.3 Lower bound

In this subsection, we will define a new lower bound for our VTSFE model to use as loss function. We will make the calculation whith the full DMP transition model to obtain a general lower bound, whether we use the point-attractor aspect or not.

Contrary to [2], we won't be using the DVBF lower bound, as the main difference in VTSFE and VAE-DMP with respect to DVBF is that unknown forcing terms f_t replace observed commands u_t. This means that we could look for a different variational lower bound using $\ln(p_\theta(x_{1:T}))$ rather than $\ln(p_\theta(x_{1:T}|u_{1:T}))$ as in equation (1.12).

Similarly to equation (1.4), we can write:

$$ELBO = \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})}[\ln(p_\theta(x_{1:T}|z_{1:T}))] - D_{KL}(q_\phi(z_{1:T}|x_{1:T}) \parallel p_\theta(z_{1:T})). \tag{2.8}$$

Instead of inserting the derivative in the state so as to get a first-order Markov assumption, we simply use the second-order Markov property. With a VAE encoder only for the first, second and last time steps, we can therefore write:

$$\begin{aligned} q_\phi(z_{1:T}|x_{1:T}) &= \frac{q_\phi(z_{1:T}, x_{1:T})}{q_\phi(x_{1:T})} \\ &= \prod_{t\in\{1,2,T\}} q_\phi(z_t|x_t) \prod_{t=3}^{T-1} q_\phi(z_t|x_{1:T}, z_{t-1}, z_{t-2}, z_T) \end{aligned} \tag{2.9}$$

Given $\forall t \in [3, T-1], z_t = g(z_{t-1}, z_{t-2}, z_T, \epsilon_{t-1}, f_{t-1})$, where g is a deterministic transition function, we can push (2.9) a little further :

$$\begin{aligned} & q_\phi(z_{1:T}|x_{1:T}) \\ &= \prod_{t\in\{1,2,T\}} q_\phi(z_t|x_t) \prod_{t=2}^{T-2} q_\phi(g(z_t, z_{t-1}, z_T, \epsilon_t, f_t)|x_{1:T}, z_t, z_{t-1}, z_T) \\ &= \prod_{t\in\{1,2,T\}} q_\phi(z_t|x_t) \prod_{t=2}^{T-2} q_\phi(\epsilon_t, f_t|x_{1:T}, z_t, z_{t-1}, z_T) \\ &= \prod_{t\in\{1,2,T\}} q_\phi(z_t|x_t) \prod_{t=2}^{T-2} \frac{q_\phi(\epsilon_t, f_t, x_{1:T}, z_t, z_{t-1}, z_T)}{q_\phi(x_{1:T}, z_t, z_{t-1}, z_T)} \\ &= \prod_{t\in\{1,2,T\}} q_\phi(z_t|x_t) \prod_{t=2}^{T-2} \frac{q_\phi(f_t|x_{1:T}) q_\phi(\epsilon_t|f_t, x_{t+1}, z_t, z_{t-1}, z_T) q_\phi(x_{1:T}, z_t, z_{t-1}, z_T)}{q_\phi(x_{1:T}, z_t, z_{t-1}, z_T)} \\ &= \prod_{t\in\{1,2,T\}} q_\phi(z_t|x_t) \prod_{t=2}^{T-2} q_\phi(\epsilon_t|f_t, x_{t+1}, z_t, z_{t-1}, z_T) q_\phi(f_t|x_{1:T}) \end{aligned} \tag{2.10}$$

In the same way:

$$\begin{aligned} p_\theta(z_{1:T}) &= \prod_{t\in\{1,2,T\}} p_\theta(z_t) \prod_{t=3}^{T-1} p_\theta(z_t|z_{t-1}, z_{t-2}, z_T) \\ &= \prod_{t\in\{1,2,T\}} p_\theta(z_t) \prod_{t=2}^{T-2} p_\theta(\epsilon_t, f_t|z_t, z_{t-1}, z_T) \\ &= \prod_{t\in\{1,2,T\}} p_\theta(z_t) \prod_{t=2}^{T-2} p_\theta(\epsilon_t) p_\theta(f_t) \end{aligned} \tag{2.11}$$

And finally:

$$p_\theta(x_{1:T}|z_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t|z_t) \tag{2.12}$$

(2.9) and (2.12) lead to:

$$\begin{aligned} &\mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})}[\ln(p_\theta(x_{1:T}|z_{1:T}))] \\ &= \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})}\left[\sum_{t=1}^{T} \ln(p_\theta(x_t|z_t))\right] \\ &= \sum_{t_1=1}^{T} \int_{z_{1:T}} \prod_{t_2\in\{1,2,T\}} q_\phi(z_{t_2}|x_{t_2}) \prod_{t_2=3}^{T-1} q_\phi(z_{t_2}|x_{1:T}, z_{t_2-1}, z_{t_2-2}, z_T) \ln(p_\theta(x_{t_1}|z_{t_1})) dz_{1:T} \\ &= \sum_{t_1\in\{1,2,T\}} \Bigg[\int_{z_{t_1}} q_\phi(z_{t_1}|x_{t_1}) \ln(p_\theta(x_{t_1}|z_{t_1})) \\ &\quad \times \Bigg[\int_{z_{[1:T]\setminus\{t_1\}}} \prod_{t_2\in\{1,2,T\}\setminus\{t_1\}} q_\phi(z_{t_2}|x_{t_2}) \\ &\quad\quad \times \prod_{t_2=3}^{T-1} q_\phi(z_{t_2}|x_{1:T}, z_{t_2-1}, z_{t_2-2}, z_T)\, dz_{[1:T]\setminus\{t_1\}}\Bigg]\, dz_{t_1}\Bigg] \\ &\quad + \sum_{t_1=3}^{T-1} \Bigg[\int_{z_{1:t_1}, z_T} \prod_{t_2\in\{1,2,T\}} q_\phi(z_{t_2}|x_{t_2}) \prod_{t_2=3}^{t_1} q_\phi(z_{t_2}|x_{1:T}, z_{t_2-1}, z_{t_2-2}, z_T) \ln(p_\theta(x_{t_1}|z_{t_1})) \\ &\quad\quad \times \Bigg[\int_{z_{[t_1+1:T-1]}} \prod_{t_2=t_1+1}^{T-1} q_\phi(z_{t_2}|x_{1:T}, z_{t_2-1}, z_{t_2-2}, z_T)\, dz_{[t_1+1:T-1]}\Bigg]\, dz_{1:t_1} dz_T\Bigg] \\ &= \sum_{t\in\{1,2,T\}} \left[\int_{z_t} q_\phi(z_t|x_t) \ln(p_\theta(x_t|z_t))\, dz_t\right] \\ &\quad + \sum_{t=3}^{T-1} \left[\int_{z_{1:t}, z_T} q_\phi(z_{1:t}, z_T|x_{1:T}) \ln(p_\theta(x_t|z_t))\, dz_{1:t} dz_T\right] \\ &= \sum_{t\in\{1,2,T\}} \mathbb{E}_{q_\phi(z_t|x_t)}[\ln(p_\theta(x_t|z_t))] + \sum_{t=3}^{T-1} \mathbb{E}_{q_\phi(z_{1:t}, z_T|x_{1:T})}[\ln(p_\theta(x_t|z_t))] \end{aligned} \tag{2.13}$$

Besides, (2.11) and (2.10) lead to:

$$
\begin{aligned}
&D_{KL}(q_\phi(z_{1:T}|x_{1:T}) \,||\, p_\theta(z_{1:T})) \\
&= \int_{z_{1:T}} q_\phi(z_{1:T}|x_{1:T}) \ln\left(\frac{q_\phi(z_{1:T}|x_{1:T})}{p_\theta(z_{1:T})}\right) dz_{1:T} \\
&= \int_{z_{1:T}} q_\phi(z_{1:T}|x_{1:T}) \ln\left(\frac{\prod_{t\in\{1,2,T\}} q_\phi(z_t|x_t) \prod_{t=2}^{T-2} q_\phi(\epsilon_t|f_t, x_{t+1}, z_t, z_{t-1}, z_T) q_\phi(f_t|x_{1:T})}{\prod_{t\in\{1,2,T\}} p_\theta(z_t) \prod_{t=2}^{T-2} p_\theta(\epsilon_t) p_\theta(f_t)}\right) dz_{1:T} \\
&= \int_{z_{1:T}} q_\phi(z_{1:T}|x_{1:T}) \left[\sum_{t\in\{1,2,T\}} \ln\left(\frac{q_\phi(z_t|x_t)}{p_\theta(z_t)}\right) + \sum_{t=2}^{T-2} \ln\left(\frac{q_\phi(\epsilon_t|f_t, x_{t+1}, z_t, z_{t-1}, z_T)}{p_\theta(\epsilon_t)}\right)\right. \\
&\qquad \left. + \sum_{t=2}^{T-2} \ln(q_\phi(f_t|x_{1:T})) - \sum_{t=2}^{T-2} \ln(p_\theta(f_t))\right] dz_{1:T} \\
&= \sum_{t\in\{1,2,T\}} \int_{z_t} q_\phi(z_t|x_t) \ln\left(\frac{q_\phi(z_t|x_t)}{p_\theta(z_t)}\right) dz_t \\
&\qquad + \sum_{t_1=2}^{T-2} \left[\int_{z_{1:t_1}, z_T} \int_{\epsilon_{t_1}} \int_{f_{t_1}} \prod_{t_2\in\{1,2,T\}} q_\phi(z_{t_2}|x_{t_2}) \prod_{t_2=3}^{t_1} q_\phi(z_{t_2}|x_{1:T}, z_{t_2-1}, z_{t_2-2}, z_{goal})\right. \\
&\qquad\quad \times q_\phi(\epsilon_{t_1}|f_{t_1}, x_{t_1+1}, z_{t_1}, z_{t_1-1}, z_T) q_\phi(f_{t_1}|x_{1:T}) \left[\ln\left(\frac{q_\phi(\epsilon_{t_1}|f_{t_1}, x_{t_1+1}, z_{t_1}, z_{t_1-1}, z_T)}{p_\theta(\epsilon_{t_1})}\right)\right. \\
&\qquad\qquad \left.\left. + \ln(q_\phi(f_{t_1}|x_{1:T})) - \ln(p_\theta(f_{t_1}))\right]\right. \\
&\qquad\quad \left.\times \left[\int_{z_{[t_1+1:T-1]}} \prod_{t_2=t_1+1}^{T-1} q_\phi(z_{t_2}|x_{1:T}, z_{t_2-1}, z_{t_2-2}, z_{goal})\, dz_{[t_1+1:T-1]}\right] df_{t_1} d\epsilon_{t_1} dz_{1:t_1} dz_T\right] \\
&= \sum_{t\in\{1,2,T\}} D_{KL}(q_\phi(z_t|x_t) \,||\, p_\theta(z_t)) \\
&\qquad + \sum_{t_1=2}^{T-2} \mathbb{E}_{q_\phi(z_{1:t_1}, z_T|x_{1:T})} \left[\int_{f_{t_1}} q_\phi(f_{t_1}|x_{1:T})\right. \\
&\qquad\quad \left[\int_{\epsilon_{t_1}} q_\phi(\epsilon_{t_1}|f_{t_1}, x_{t_1+1}, z_{t_1}, z_{t_1-1}, z_T)\, \ln\left(\frac{q_\phi(\epsilon_{t_1}|f_{t_1}, x_{t_1+1}, z_{t_1}, z_{t_1-1}, z_T)}{p_\theta(\epsilon_{t_1})}\right) d\epsilon_{t_1}\right. \\
&\qquad\quad + \ln(q_\phi(f_{t_1}|x_{1:T})) \int_{\epsilon_{t_1}} q_\phi(\epsilon_{t_1}|f_{t_1}, x_{t_1+1}, z_{t_1}, z_{t_1-1}, z_T)\, d\epsilon_{t_1} \\
&\qquad\quad \left.\left. - \ln(p_\theta(f_{t_1})) \int_{\epsilon_{t_1}} q_\phi(\epsilon_{t_1}|f_{t_1}, x_{t_1+1}, z_{t_1}, z_{t_1-1}, z_T) d\epsilon_{t_1}\right] df_{t_1}\right]
\end{aligned}
\tag{2.14}
$$

As f_t is inferred deterministically by $x_{1:T}$, the probability of f_t given $x_{1:T}$ equals 1. Thus, the distribution $q_\phi(f_t|x_{1:T})$ is a Dirac function centered at $\hat{f}_t$, where $\hat{f}_t$ is the inferred value of f_t for a given $x_{1:T}$. Therefore, $\int_{f_{t_1}} q_\phi(f_{t_1}|x_{1:T})\, \ln(q_\phi(f_{t_1}|x_{1:T}))\, df_{t_1} = \ln(1) = 0$ and $\int_{f_{t_1}} q_\phi(f_{t_1}|x_{1:T})\, \ln(p_\theta(f_{t_1}))\, df_{t_1} =$

$\ln(p_\theta(\hat{f}_t))$. Thereby, (2.14) simplifies as following:

$$\begin{aligned}
&D_{KL}(q_\phi(z_{1:T}|x_{1:T}) \,||\, p_\theta(z_{1:T})) \\
&= \sum_{t\in\{1,2,T\}} D_{KL}(q_\phi(z_t|x_t) \,||\, p_\theta(z_t)) \\
&\quad + \sum_{t_1=2}^{T-2} \mathbb{E}_{q_\phi(z_{1:t_1},z_T|x_{1:T})} \Bigg[-\ln(p_\theta(\hat{f}_{t_1})) \\
&\qquad + \int_{\epsilon_{t_1}} q_\phi(\epsilon_{t_1}|\hat{f}_{t_1}, x_{t_1+1}, z_{t_1}, z_{t_1-1}, z_T) \, \ln\left(\frac{q_\phi(\epsilon_{t_1}|\hat{f}_{t_1}, x_{t_1+1}, z_{t_1}, z_{t_1-1}, z_T)}{p_\theta(\epsilon_{t_1})}\right) d\epsilon_{t_1} \Bigg] \\
&= \sum_{t\in\{1,2,T\}} D_{KL}(q_\phi(z_t|x_t) \,||\, p_\theta(z_t)) \\
&\quad + \sum_{t=2}^{T-2} \mathbb{E}_{q_\phi(z_{1:t},z_T|x_{1:T})} \Big[-\ln(p_\theta(\hat{f}_t)) \\
&\qquad + D_{KL}(q_\phi(\epsilon_t|\hat{f}_t, x_{t+1}, z_t, z_{t-1}, z_T) \,||\, p_\theta(\epsilon_t)) \Big]
\end{aligned} \tag{2.15}$$

Then, following [3] and using (2.3), we get:

$$\begin{aligned}
&D_{KL}(q_\phi(\epsilon_t|f_t, x_{t+1}, z_t, z_{t-1}, z_T) \,||\, p_\theta(\epsilon_t)) \\
&= \frac{1}{2}\left[\ln\left(\frac{\left|\sigma^2_{scale} I\right|}{\left|\sigma^2_{scale} I \sigma^2_{\epsilon,t} I\right|}\right) - d_z + \operatorname{tr}\left((\sigma^2_{scale} I)^{-1}\sigma^2_{scale} I \sigma^2_{\epsilon,t} I\right) + (-\sigma_{scale} \odot \mu_{\epsilon,t})(\sigma^2_{scale} I)^{-1}(-\sigma_{scale} \odot \mu_{\epsilon,t})^T\right] \\
&= \frac{1}{2}\sum_{i=1}^{d_z}\left[-\ln(\sigma^2_{\epsilon,t,i}) - 1 + \sigma^2_{\epsilon,t,i} + \mu^2_{\epsilon,t,i}\right]
\end{aligned} \tag{2.16}$$

Using the same discrete approximation of the mean as in (1.5), this leads us to a computationnaly highly expensive approximation in $\Theta(L^{T-3})$. In order to be able to compute that new lower bound, we made a rougher approximation by taking the L^3 samples from the three priors $p_\theta(z_1)$, $p_\theta(z_2)$ and $p_\theta(z_T)$, and for all $t \in [2, T-2]$, only propagating $z_{t+1} = g(\mu_{z,T}, \mu_{z,t}, \mu_{z,t-1}, \mu_{\epsilon,t}, \hat{f}_t)$ for each of these L^3 3-tuples $(z_{l_1}, z_{l_2}, z_{l_T})$, which changes (2.3) to:

$$\begin{aligned}
&q_\phi(\epsilon_t|f_t, x_{t+1}, z_t, z_{t-1}, z_T) \\
&\quad = \mathcal{N}(\sigma_{scale} \odot \mu_{\epsilon,t},\ \sigma^2_{scale} I \sigma^2_{\epsilon,t} I + K^2 \sum_{t'=2}^{t-1} \sigma^2_{scale} I \sigma^2_{\epsilon,t'} I) \\
&\quad = \mathcal{N}(\sigma_{scale} \odot \mu_{\epsilon,t},\ \sigma^2_{scale} I \left[\sigma^2_{\epsilon,t} + K^2 \sum_{t'=2}^{t-1} \sigma^2_{\epsilon,t'}\right] I) \\
&p_\theta(\epsilon_t) = \mathcal{N}(0, [1 + (t-2)K^2]\sigma^2_{scale} I)
\end{aligned} \tag{2.17}$$

as g is a linear transformation. K is the real factor that multiplies ϵ_t in g.

Thereby, (2.16) becomes:

$$\begin{aligned}
&D_{KL}(q_\phi(\epsilon_t|f_t, x_{t+1}, z_t, z_{t-1}, z_T) \,||\, p_\theta(\epsilon_t)) \\
&= \frac{1}{2}\sum_{i=1}^{d_z}\Bigg[\ln(1 + (t-2)K^2) - \ln(\sigma^2_{\epsilon,t,i} + K^2 \sum_{t'=2}^{t-1} \sigma^2_{\epsilon,t',i}) \\
&\quad -1 + \frac{\sigma^2_{\epsilon,t,i} + K^2 \sum_{t'=2}^{t-1} \sigma^2_{\epsilon,t',i}}{1 + (t-2)K^2} + \frac{\mu^2_{\epsilon,t,i}}{1 + (t-2)K^2}\Bigg]
\end{aligned} \tag{2.18}$$

Then we obtain, $\forall t \in [2, T-2]$:

$$\begin{aligned}
&\mathbb{E}_{q_\phi(z_{1:t}, z_T|x_{1:T})}[D_{KL}(q_\phi(\epsilon_t|f_t, x_{t+1}, z_t, z_{t-1}, z_T) \mid\mid p_\theta(\epsilon_t))] \\
&\approx \frac{1}{L^3} \sum_{z_{l_1}=1}^{L} \sum_{z_{l_2}=1}^{L} \sum_{z_{l_T}=1}^{L} \frac{1}{2} \sum_{i=1}^{d_z} \Bigg[\ln(1 + (t-2)K^2) \\
&\quad - \ln(\sigma_{\epsilon,t,i}^2 + K^2 \sum_{t'=2}^{t-1} \sigma_{\epsilon,t',i}^2) - 1 \\
&\quad + \frac{\mu_{\epsilon,t,i}^2 + \sigma_{\epsilon,t,i}^2 + K^2 \sum_{t'=2}^{t-1} \sigma_{\epsilon,t',i}^2}{1 + (t-2)K^2} \Bigg]_{z_{l_1}, z_{l_2}, z_{l_T}}
\end{aligned} \tag{2.19}$$

which leaves us in the end with $\sum_{t=2}^{T-2} L^3$ samples. That approximation has also been used on 2.13 with 4-tuples $(z_{l_1}, z_{l_2}, z_{l_t}, z_{l_T})$ instead of the previous 3-tuples in order to sample the latent space at each inference on z as in VAE, leading to $\Theta(L^4)$, i.e. $\forall t \in [2, T-2]$:

$$\begin{aligned}
&- \mathbb{E}_{q_\phi(z_{[1:t]}, z_T|x_{1:t})}[ln(p_\theta(x_t|z_t))] \\
&\approx \frac{1}{L^4} \sum_{z_{l_1}=1}^{L} \sum_{z_{l_2}=1}^{L} \sum_{z_{l_t}=1}^{L} \sum_{z_{l_T}=1}^{L} \frac{1}{2} \sum_{i=1}^{d_x} \Bigg[ln(2\pi) \\
&\quad + ln(\sigma_{x,t,i}^2) \\
&\quad + \frac{(x_{t,i} - \mu_{x,t,i})^2}{\sigma_{x,t,i}^2} \Bigg]_{z_{l_1}, z_{l_2}, z_{l_t}, z_{l_T}}
\end{aligned} \tag{2.20}$$

$\forall t \in \{1, 2, T\}$, (1.5) is used. Nevertheless, it's worth mentioning that $\sigma_{x,t}$ is only used in the reconstruction loss term, without any other constraint. Therefore, it acts as a degree of freedom at optimization time that can compensate errors either by being high to reduce the squared error or by being very close to zero to make the log take a negative value if the error is small, which in particular could happen if the input data range doesn't vary much around zero, as it is the case for our data that ranges in $[-1, 1]$. Besides, having a different $\sigma_{x,t}$ for each z sample causes the optimization process to treat each error differently which reinforces these issues. Thus, in order to correct that error compensation and to save gradient computations, we made an additional assumption on $p_\theta(x_t|z_t)$ by setting $\sigma_{x,t}^2 = \frac{1}{2\pi}$, where the value $\frac{1}{2\pi}$ has been chosen to remove all constant term so the reconstruction error is entirely due to the squared term. That new assumption then allows us to replace the expression of $\ln(p_\theta(x_t|z_t))$ in both (1.5) and (2.20) by:

$$\ln(p_\theta(x_t|z_t)) = \pi(x_{t,i} - \mu_{x,t,i})^2 \tag{2.21}$$

Let's now define the last term of our lower bound. Unfortunately, $p_\theta(\hat{f}_t)$ is intractable. We could simply continue without it though, giving us a lower bound of $D_{KL}(q_\phi(z_{1:T}|x_{1:T}) \mid\mid p_\theta(z_{1:T}))$. Nevertheless, interpreting it as a loss function term, we notice that, being the result of a Kullback-Leibler divergence with a Dirac distribution, it acts as a penalty if $P_\theta(\hat{f}_t) \neq 1$. Therefore, considering it as a loss term on dynamics reconstruction, we want to keep at least a part of that term. Given the fact that without any knowledge, f_t could be equally anything, as x_t, $p_\theta(f_t)$ should be closer to a uniform distribution over an infinite space than a Dirac distribution centered on a particular $\hat{f}$. Thus, we assume that $-\mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})}[\ln(p_\theta(\hat{f}_t))] \geq -\mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})}[\ln(p_\theta(\hat{f}_t|z_{1:T}))] \geq 0$ since $\hat{f}_t$ is used to infer $z_{1:T}$.

So we need to define $p_\theta(f_t|z_{1:T})$. First, we make the assumption that it follows a gaussian distribution. Then, given the deterministic transition function $z_{t+1} = g(z_{l_T}, z_t, z_{t-1}, \epsilon_t, f_t)$, where ϵ_t is simply added to f_t, we notice that the only variance contained in $p_\theta(f_t|z_{1:T})$ is the one of $p_\theta(\epsilon_t)$. Besides:

$$\begin{aligned} p_\theta(f_t|z_{1:T}) &= \int_{x_{1:T}} p_\theta(f_t|x_{1:T}, z_{1:T}) p_\theta(x_{1:T}|z_{1:T})\ dx_{1:T} \\ &\approx \int_{x_{1:T}} q_\phi(f_t|x_{1:T}) p_\theta(x_{1:T}|z_{1:T})\ dx_{1:T}. \end{aligned} \tag{2.22}$$

Thereby, we make the assumption that $p_\theta(f_t|z_{1:T}) \approx \mathcal{N}(d(x_{1:T})_t, [1 + (t-2)K^2]\sigma^2_{scale} I)$, where $x_{1:T} \sim p_\theta(x_{1:T}|z_{1:T})$ and d represents a function taking $x_{1:T}$ as input and $f_{1:T}$ as output where $f_{1:T} \sim q_\phi(f_t|x_{1:T})$, as the MLP used to infer $f_{1:T}$.

Unfortunately, even combined with the previously defined $\Theta(L^3)$ approximation of $\mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})}$, the additional sampling on $x_{1:T}$ still leads to a $\Theta(L^3 \times X^T)$ approximation, where X is the number of samples taken for x_t. Therefore, we made an even rougher approximation by taking $X = 1$, where that only sample is the T-tuple $(\mu_{x,t})_{t\in[1:T]}$ and $\mu_{x,t}$ is the mean of $p_\theta(x_t|z_t)$. Thus, we have, $\forall t \in [2, T-2]$:

$$\begin{aligned} &-\mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})}[ln(p_\theta(\hat{f}_t|z_{1:T}))] \\ &\quad\approx \frac{1}{L^3} \sum_{z_{l_1}=1}^{L} \sum_{z_{l_2}=1}^{L} \sum_{z_{l_T}=1}^{L} \frac{1}{2} \sum_{i=1}^{d_z} \Bigg[ln(2\pi) \\ &\qquad + ln([1 + (t-2)K^2]\sigma^2_{scale,i}) \\ &\qquad + \frac{(\hat{f}_{t,i} - d(\mu_{x,1:T,i})_t)^2}{[1 + (t-2)K^2]\sigma^2_{scale,i}} \Bigg]_{z_{l_1}, z_{l_2}, z_{l_T}} \end{aligned} \tag{2.23}$$

Finally, we obtain the following Evidence Lower BOunds (*ELBO*):

$$ELBO \leq ELBO_{VTSFE} \leq ELBO_{VAE\text{-}DMP} \tag{2.24}$$

where:

$$
\begin{aligned}
&ELBO_{VTSFE} \\
&= \Bigg(\sum_{t \in \{1,2,T\}} \mathbb{E}_{q_\phi(z_t|x_t)}[\ln(p_\theta(x_t|z_t))] \\
&\quad + \sum_{t=3}^{T-1} \Big[\mathbb{E}_{q_\phi(z_{1:t}, z_T|x_{1:T})}[\ln(p_\theta(x_t|z_t))] \\
&\quad + \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})}[\ln(p_\theta(\hat{f}_{t-1}|z_{1:T}))] \Big] \Bigg) \\
&\quad - \Bigg(\sum_{t \in \{1,2,T\}} D_{KL}(q_\phi(z_t|x_t) \,||\, p_\theta(z_t)) \\
&\quad + \sum_{t=2}^{T-2} \mathbb{E}_{q_\phi(z_{1:t}, z_T|x_{1:T})} \left[D_{KL}(q_\phi(\epsilon_t|\hat{f}_t, x_{t+1}, z_t, z_{t-1}, z_T) \,||\, p_\theta(\epsilon_t)) \right] \Bigg) \\
&\leq \Bigg(\sum_{t \in \{1,2,T\}} \mathbb{E}_{q_\phi(z_t|x_t)}[\ln(p_\theta(x_t|z_t))] \\
&\quad + \sum_{t=3}^{T-1} \mathbb{E}_{q_\phi(z_{1:t}, z_T|x_{1:T})}[\ln(p_\theta(x_t|z_t))] \Bigg) \\
&\quad - \Bigg(\sum_{t \in \{1,2,T\}} D_{KL}(q_\phi(z_t|x_t) \,||\, p_\theta(z_t)) \\
&\quad + \sum_{t=2}^{T-2} \mathbb{E}_{q_\phi(z_{1:t}, z_T|x_{1:T})} \left[D_{KL}(q_\phi(\epsilon_t|\hat{f}_t, x_{t+1}, z_t, z_{t-1}, z_T) \,||\, p_\theta(\epsilon_t)) \right] \Bigg) \\
&\approx \Bigg(\sum_{t=1}^{T} \mathbb{E}_{q_\phi(z_t|x_t)}[\ln(p_\theta(x_t|z_t))] \Bigg) \\
&\quad - \Bigg(\sum_{t \in \{1,2,T\}} D_{KL}(q_\phi(z_t|x_t) \,||\, p_\theta(z_t)) \\
&\quad + \sum_{t=3}^{T-1} D_{KL}(q_\phi(\epsilon_t|\hat{f}_t, x_{t+1}, z_t, z_{t-1}, z_T) \,||\, p_\theta(\epsilon_t)) \Bigg)
\end{aligned}
\tag{2.25}
$$

where the latter expression is equivalent to the lower bound of VAE-DMP: $ELBO_{VAE\text{-}DMP}$.

More precise mean approximations aside, our lower bound $ELBO_{VTSFE}$ is thereby tighter to the initial *ELBO* than the DVBF one from which derives $ELBO_{VAE\text{-}DMP}$ thanks to our *dynamics reconstruction loss term* $-\mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})}[\ln(p_\theta(\hat{f}_t|z_{1:T}))]$.

However, due to our approximations, we end up with a complexity, given T, of $\Theta(L^4)$ (or more precisely of $\Theta(P^3 \times M)$, where P is the number of samples from each of the priors $(z_{l_1}, z_{l_2}, z_{l_T})$ to propagate through time, and M is the number of samples from ϵ_{t-1} at time t).

That new lower bound is more computationally expensive, but should allow the creation of a more generalizable latent space by creating local space continuity in terms of both surroundings of each input and trajectory of the whole sequence, i.e. reducing the intraclass variance in latent

space for each movement type, for L^3 trajectory initialization samples (prior sampling). At the same time, it forces the reconstructed inputs $x_{1:T}$ to keep the same dynamics as the actual inputs, in addition to force them to be individually close to the actual ones.

To reduce the complexity, we propose a *light VTSFE* by making a rough $\Theta(L)$ approximation consisting of only one prior sampling $(\mu_{z,1}, \mu_{z,2}, \mu_{z,T})$ for D_{KL} and only L samples $(\mu_{z,1}, \mu_{z,2}, \mu_{z,T}, z_{l_t})$ for $-\mathbb{E}_{q_\phi(z_{[1:t]}, z_T|x_{1:t})}[ln(p_\theta(x_t|z_t))]$. This version reduces intraclass variance in latent space for each movement type, but only processes one trajectory initialization sample (prior sampling).

Bibliography

[1] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *eprint arXiv:1601.00670*, 2016. 1

[2] Nutan Chen, Maximilian Karl, and Patrick Van Der Smagt. Dynamic movement primitives in latent space of time-dependent variational autoencoders. *16th IEEE-RAS International Conference on Humanoid Robots*, 2016. 4, 5, 7, 9

[3] John Duchi. Derivations for linear algebra and optimization. http://web.stanford.edu/~jduchi/projects/general_notes.pdf. 12

[4] Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick Van Der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. *The International Conference on Learning Representations (ICLR)*, 2017. 3

[5] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *The International Conference on Learning Representations (ICLR)*, 2014. 1