

Dynamical Variational Autoencoders :  
discrete-time and continuous-time models.  
Links to Stochastic calculus

ENS Paris-Saclay, MVA

Benjamin Deporte : [benjamin.deporte@ens-paris-saclay.fr](mailto:benjamin.deporte@ens-paris-saclay.fr)

August 2025

## Abstract

This report describes a particular class of Variational Auto Encoders (VAEs) : the Dynamical Variational Auto Encoders (DVAEs). DVAEs are a specific class of models, adapted to the study of data sequences. In DVAEs, the latent variables are structured themselves as a correlated set (usually a sequence also), aiming at encoding the temporal dimension of the data. We will review the general formulation of DVAEs and the detailed implementation of three models : extended Kalman filter, Variational Recurrent Neural Network (VRNN), and Gaussian Process Variational Auto Encoder (GP-VAE).

We then provide an overview of the information theory framework for data sequences -specifically some results on entropy rates- as an attempt to empirically quantify the degree of "randomness" of data sequences. In doing so, we can try and evaluate the expressiveness of DVAEs, and their relative performance with respect to other well-known models (such as Long Short Term Memorys (LSTMs)).

However, it is the theory of stochastic calculus that will provide us with great insights regarding the expressiveness of DVAEs. First, we will see that the solution of a Stochastic Differential Equation (SDE) is a Markov Process, that can quite naturally be expressed as a DVAE. Furthermore, if the SDE is actually linear, we will see that its solution is actually a Gaussian Process, and fits naturally into the GP-VAE model. All the associated results of the kernels theory (notably results on the spectral theory) then apply.

Armed with those results, we carry out some experiments, to demonstrate... SURPRISE.

The code is available at : [https://github.com/BenjaminDeporte/MVA\\_Stage](https://github.com/BenjaminDeporte/MVA_Stage)

## **Acknowledgements**

# Contents

<b>I</b>	<b>Introduction</b>	<b>7</b>
1	Subject	8
2	Report Outline	9
3	Related work	10
<b>II</b>	<b>Dynamical Variational Autoencoders</b>	<b>11</b>
4	D-separation	13
5	Dynamical Variational Auto Encoders	16
6	Deep Kalman Filter	20
7	Variational Recurrent Neural Network	23
8	Gaussian Process Variational Auto Encoder	26
<b>III</b>	<b>Notions on stochastic differential equations and their relationships to DVAEs</b>	<b>31</b>
9	Stochastic processes basics	33
10	Stochastic Calculus	41
11	Ito's calculus and SDE	46
11.1	Stochastic differential equations - SDE . . . . .	47
12	Quantifying randomness with Entropy	52
13	Perspective 1 : Linear SDEs and Markovian kernel GPs	56
14	Perspective 2 : more IT on stochastic processes	57
<b>IV</b>	<b>Experiments</b>	<b>58</b>
15	Experiments	59

<b>V</b>	<b>Conclusion and Discussion</b>	<b>60</b>
<b>16</b>	<b>Conclusions</b>	<b>61</b>
<b>VI</b>	<b>Appendices</b>	<b>62</b>
<b>Appendices</b>		<b>63</b>
<b>A</b>	<b>Vanilla Variational Auto Encoder</b>	<b>64</b>
<b>B</b>	<b>Gaussian Process</b>	<b>66</b>
<b>C</b>	<b>KL divergence between two exponential-family distributions</b>	<b>68</b>

# List of Figures

6.1	Probabilistic model of a Deep Kalman Filter . . . . .	20
6.2	Deep Kalman Filter Model Architecture . . . . .	22
7.1	Probabilistic model of a Variational RNN . . . . .	23
7.2	Variational RNN Model Architecture . . . . .	25
8.1	Probabilistic model of a GP-VAE . . . . .	27
8.2	Gaussian Process VAE Model Architecture . . . . .	30
A.1	Vanilla VAE . . . . .	64

## **Part I**

# **Introduction**

Variational Autoencoders are a well-known class of generative models, where the latent variables are usually assumed to be independent and identically distributed. This assumption is inappropriate when the data is time-dependent, such as in time-series, images sequences or videos. It is then natural to structure some sort of temporal dependency in the latent prior : this is the main idea of Dynamical VAEs.

A first question is whether the use of Dynamical VAEs on time-dependent data allow better performance than the "legacy" models. A natural test framework is time series, where the usual models -such as ARIMA- have been performing successfully for quite a while.

A second question takes the thinking a little bit deeper. If we consider a time-dependent data sequence as the realization of a stochastic process -which is quite natural for time series, but can be envisioned as well for videos or other sources-, then we infer that some data sequences are, by design, easier than others to predict and learn generative models on. An idea is then to quantify the "randomness" of a data sequence -and more importantly, the "randomness" of the underlying stochastic process-, so we can measure when the use of Dynamical VAEs is likely to add value. A natural tool is Information Theory, where some results exist regarding stochastic processes.

Last but not least, assuming that we have learnt "reasonably well" a generative model, we can wonder how good the model performs at detecting anomalies in the data sequence. The Pandora's box of anomaly detection is vast, and covers connex notions such as detecting outliers in a stationary distribution, detecting shifts of a distribution towards a new one, etc. If using the likelihood of the data according to the learnt distribution is quite straightforward an idea, some recent results show that this may be misleading.



**Part II** covers the mathematical background that is required for most of the report. Section 1 covers the main results regarding stochastic processes and introduces the Brownian motion. (Stochastic Differential Equations are introduced later in Part III). Section 2 recaps the main first definitions and results regarding information theory, its application to stochastic processes, some theoretical results for stationary processes and experimental metrics.

**Part II** covers discrete-time dynamical Variational AutoEncoders. Section 1 is a recap of graphical models and D-Separation, that is used throughout the report. Section 2 introduces the generic formulation of DVAEs : generative and inference models, variational lower bound. Section 3 presents the first, and most simple model : the Deep Kalman Filter. Section 4 presents the most expressive model, ie the Variational RNN. Section 5 is a summary of the take-aways regarding the PyTorch-implementations.

**Part III** takes us to continuous-time Dynamical VAEs. Section 1 is a recap of the theory regarding Gaussian Processes, that will be at the core of this part. Section 2 is intended to be a self-sustained presentation of the stochastic calculus, where we go quickly over the construction of the Ito's integral to derive the necessary Ito's lemma and subsequent calculation rules for the rest of the report. Section 3 presents the GP-VAE model. Section 4 is a summary of the take-aways regarding the PyTorch implementation.

**Part IV** summarizes the experiments : Blembet UK tides, Brownian motion, O.U + Student observation model, cryptos...

**Part V** is an opening to what could be covered next ! Section 1 presents the link between SDEs and Markovian kernel Gaussian Processes. Section 2 presents some results regarding information theory and stochastic processes.

## Related work

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

## **Part II**

# **Dynamical Variational Autoencoders**

This part presents the general framework of DVAEs. We start with a reminder of the key notion of **D-separation**, which is central in Graphical Probabilistic Models (GPMs).

Then, we describe three models in details:

- **Deep Kalman Filter** : this model arises as the first evolution of the well-known Kalman Filter, with richer, Multi Layer Perceptrons (MLPs) networks for the encoder and decoder.
- **Variational Recurrent Neural Network** : at the other end of the spectrum, VRNNs provide the most expressive discrete-time formulation of the encoder and decoder. We describe here a different implementation from the one described in [14].
- **Gaussian Process Variational Auto Encoder** : in this model, the prior over the latent variables is no longer discrete, but is a Gaussian Process (GP). This allows for sampling data at irregular intervals. Another benefit is the use of richer kernel families to encode prior knowledge.

## D-separation

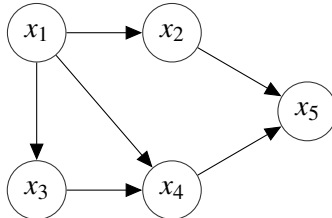
Graphical models are an efficient way to describe families of factorized joint distributions of a data set  $(x_i)_{i=1,\dots,n}$  into a Directed Acyclic Graph (DAG).

Given such a dataset, we can build a DAG where each node is indexed by an integer that is higher than the indexes of its *parent* nodes, such that the joint distribution over the dataset factorizes as:

$$p(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p(x_i | pa_i) \quad (4.1)$$

where  $pa_i$  is the set of parent nodes of  $x_i$ .

For example, the following DAG



describes:

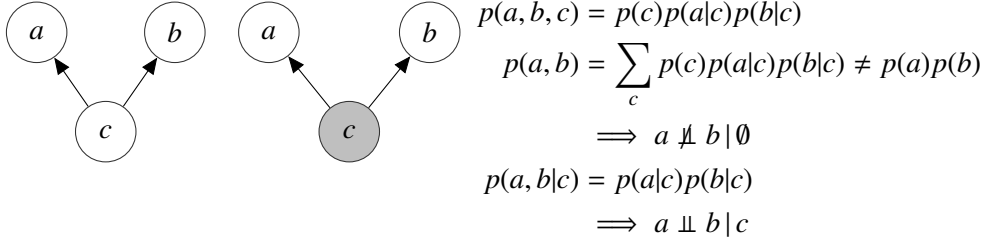
$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_1, x_3)p(x_5|x_2, x_4) \quad (4.2)$$

Describing a factorized joint probability distribution by a DAG allows to determine graphically whether two sets of nodes (ie random variables) are independent, conditioned on a third set of nodes. This allows subsequently to simplify the expressions of the observation models (ie  $p(x|z)$ , and/or the posterior models ( $q(z|x)$ ).

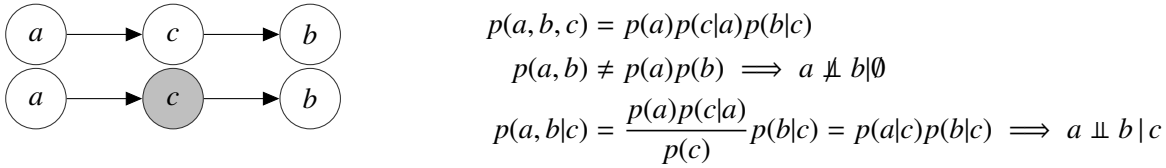
**D-Separation** is the set of rules that determine whether there is conditional independence between two sets given a third one. D-Separation is well described in key books such as [3], [17] or [15]. We will enunciate here the key concepts, and refer the interested reader to those books.

D-Separation is a way to find out graphically conditional (in)dependence relationships between random variables, that would be more difficult to calculate by marginalizing the joint distribution over the conditioning variables. A nice way to demonstrate this, is to review the three examples of 3-node DAG. NB : The observed (ie conditioning) variables are noted with gray background.

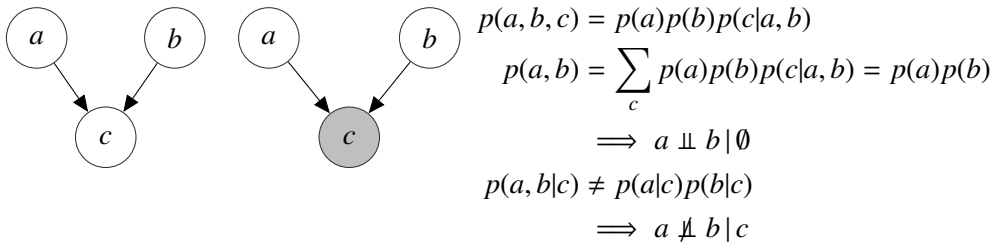
**Example 1 :**  $c$  is said *tail-to-tail* with  $a$  and  $b$ , and *blocks the path between  $a$  and  $b$* , making them conditionally independent :  $a \not\perp b | \emptyset, a \perp b | c$



**Example 2 :**  $a, c, b$  form a Markov chain.  $c$  is said *head-to-tail* with  $a$  and  $b$  and, here also, *blocks the path between  $a$  and  $b$* , making them conditionally independent :  $a \not\perp b | \emptyset, a \perp b | c$



**Example 3 :**  $c$  is said *head-to-head* with  $a$  and  $b$ . In this head-to-head configuration, contrary to the two examples above, *the path between  $a$  and  $b$  is blocked when  $c$  is unobserved* :  $a \perp b | \emptyset, a \not\perp b | c$



We can extend the notion to full sets of nodes.

## D-Separation

Let  $\mathcal{G}$  be a DAG.

Let  $A, B, C$  three disjoint sets of nodes in  $\mathcal{G} : A \cap B = A \cap C = B \cap C = \emptyset$ .

$C$  is the set of "conditioning nodes", or "observed variables".

We aim to determine whether  $A \perp\!\!\!\perp B | C$ .

### Algorithm

#### 1. Evaluate each path between $A$ and $B$

Evaluate each possible path between any point  $a \in A$ , and any point  $b \in B$ . Such a path between  $a$  and  $b$  is said **blocked** if it contains one node  $n$  such that one of two following conditions is true:

- arrows in the path are *head-to-tail* or *tail-to-tail* at node  $n$ , and  $n \in C$  ( $n$  is an observed/conditioning node).
- arrows in the path are *head-to-head* at node  $n$ , and  $n \notin C$  and none of  $n$  descendants is in  $C$

#### 2. Assess all paths

- If all paths  $(a, b), a \in A, b \in B$  are blocked, then  $A$  is said **D-separated** from  $B$  by  $C$ , and the joint distribution defined by  $\mathcal{G}$  verifies  $A \perp\!\!\!\perp B | C$ .
- If there is at least one path  $(a, b), a \in A, b \in B$  that is not blocked then  $A \not\perp\!\!\!\perp B | C$ .

## Dynamical Variational Auto Encoders

VAE models are well known and documented (see for example the seminal paper [11]). (A self-contained brief summary of VAE can be found in appendix A).

When dealing with sequential data, the i.i.d assumption on latent variables  $z_i$  is a limitation. By D-separation, all  $x_i$ 's are independent of each other conditionally by  $z_i$  :  $p(x_i|x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n, z_i) = p(x_i|z_i)$ . Therefore, a vanilla VAE can not account for correlations between  $x_i$  across time.

DVAEs encode a temporal dependency in the latent variables prior distribution. In this chapter, we review the general discrete-time setting, where the latent variables are countable and indexed by time. An exhaustive review of discrete-time DVAEs can be found in [14].

We start by some notations.

### Notations

- the data is a sequence of  $T$  points noted  $x_{1:T} = \{(x_t)_{t=1, \dots, T}\} \in \mathbb{R}^F$ .
- the sequence of the associated  $T$  latent variables is  $z_{1:T} = \{(z_t)_{t=1, \dots, T}\} \in \mathbb{R}^L$
- optionally, there may be a sequence of -usually deterministic-  $T$  inputs  $u_{1:T} = \{(u_t)_{t=1, \dots, T}\} \in \mathbb{R}^U$

The generative model is given by the general expression of the joint distribution (here with a sequence of inputs)  $p(x_{1:T}, z_{1:T}|u_{1:T})$ :

$$\begin{aligned}
 p(x_{1:T}, z_{1:T}|u_{1:T}) &= \prod_{t=1}^T p(x_t, z_t|x_{1:t-1}, z_{1:t-1}, u_{1:T}) \\
 &= \prod_{t=1}^T p(x_t|x_{1:t-1}, z_{1:t}, u_{1:T})p(z_t|x_{1:t-1}, z_{1:t-1}, u_{1:T}) \\
 &= \prod_{t=1}^T p(x_t|x_{1:t-1}, z_{1:t}, u_{1:t})p(z_t|x_{1:t-1}, z_{1:t-1}, u_{1:t})
 \end{aligned}$$

where the only assumption that is made is a causal dependency of the  $x_t, z_t$  on the inputs  $u_{1:t}$ , thus allowing to change the conditioning  $|u_{1:T}$  into  $|u_{1:t}$

In the rest of the report, we will consider systems with no input, and drop the conditioning on  $u_{1:t}$  to simplify



notations. However, the reasoning remains the same with inputs.

The true posterior  $p(z_{1:T}|x_{1:T})$  is usually untractable, but can be developed:

$$p(z_{1:T}|x_{1:T}) = \prod_{t=1}^T p(z_t|z_{1:t-1}, x_{1:T})$$

It can be noted that the true posterior exhibits a dependence of  $z_t$  on *past*  $z_{1:t-1}$ , but a dependence on the *whole* data sequence  $x_{1:T}$  (think Kalman smoother).

As in vanilla VAEs, the inference model is the approximation of the true posterior by an parametric encoder  $q_\phi(z_{1:T}|x_{1:T})$ , where  $\phi$  is the set of parameters:

$$q_\phi(z_{1:T}|x_{1:T}) = \prod_{t=1}^T q_\phi(z_t|z_{1:t-1}, x_{1:T})$$

Depending on the chosen graphical models and the corresponding D-separation results, the observation model  $p_{\theta_x}(x_t|x_{1:t-1}, z_{1:t}, u_{1:t})$  (with  $\theta_x$  the set of parameters of the observation model) and approximate posterior  $q_\phi(z_t|z_{1:t-1}, x_{1:T})$  will simplify.

It is also considered a good practice to copy the expression of  $q_\phi(z_t|z_{1:t-1}, x_{1:T})$  from the expression of the true posterior resulting from the D-separation analysis (see next chapters for examples).

Equipped with the generative model and the inference model, we compute the log likelihood of the data  $x_{1:T}$  and derive an Variational Lower Bound (VLB) for training (using the same manipulation as for vanilla VAE : multiplying both sides of the equation by  $q_\phi$  and integrating over  $dz_{1:T}$ )

$$\log p(x_{1:T}) = \log \frac{p(x_{1:T}, z_{1:T})}{p(z_{1:T}|x_{1:T})} \quad (5.1)$$

$$= \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})} \log \frac{p(x_{1:T}, z_{1:T})}{q_\phi(z_{1:T}|x_{1:T})} \frac{q_\phi(z_{1:T}|x_{1:T})}{p(z_{1:T}|x_{1:T})} \quad (5.2)$$

$$= \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})} \log \frac{p(x_{1:T}, z_{1:T})}{q_\phi(z_{1:T}|x_{1:T})} + \mathbb{KL}(q_\phi(z_{1:T}|x_{1:T}) || p(z_{1:T}|x_{1:T})) \quad (5.3)$$

$$\geq \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})} \log \frac{p(x_{1:T}, z_{1:T})}{q_\phi(z_{1:T}|x_{1:T})} = \mathcal{L}(\theta, \phi, X) \quad (5.4)$$

The dependence of  $\mathcal{L}(\theta, \phi, X)$  on  $\theta$  is made more obvious when developing  $\mathcal{L}(\theta, \phi, X)$ .

Remember we have (making the set of parameters explicit) :

$$p_\theta(x_{1:T}, z_{1:T}) = \prod_{t=1}^T p_{\theta_x}(x_t|x_{1:t-1}, z_{1:t}) p_{\theta_z}(z_t|z_{1:t-1}, x_{1:t-1}) \quad (5.5)$$

$$q_\phi(z_{1:T}|x_{1:T}) = \prod_{t=1}^T q_\phi(z_t|z_{1:t-1}, x_{1:T}) \quad (5.6)$$

Therefore

$$\mathcal{L}(\theta, \phi, X) = \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})} \log \left( \frac{\prod_{t=1}^T p_{\theta_x}(x_t|x_{1:t-1}, z_{1:t}) p_{\theta_z}(z_t|z_{1:t-1}, x_{1:t-1})}{\prod_{t=1}^T q_\phi(z_t|z_{1:t-1}, x_{1:T})} \right) \quad (5.7)$$

$$= \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})} \left( \sum_{t=1}^T \log p_{\theta_x}(x_t|x_{1:t-1}, z_{1:t}) - \sum_{t=1}^T \log \frac{q_\phi(z_t|z_{1:t-1}, x_{1:T})}{p_{\theta_z}(z_t|z_{1:t-1}, x_{1:t-1})} \right) \quad (5.8)$$

At this point, the expectations require some work. First, we note that, as  $q_\phi$  develops as 5.6, for any function  $\Psi$ , the first expectation can be written (note the change in indexes of  $z$ )

$$\mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})} \Psi(z_{1:t}) = \mathbb{E}_{q_\phi(z_{1:t}|x_{1:T})} \Psi(z_{1:t})$$

Second, we develop further and write:

$$\begin{aligned} \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})} \Psi(z_{1:t}) &= \mathbb{E}_{q_\phi(z_{1:t}|x_{1:T})} \Psi(z_{1:t}) \\ &= \mathbb{E}_{q_\phi(z_{1:t-1}|x_{1:T})} \mathbb{E}_{q_\phi(z_t|z_{1:t-1}, x_{1:T})} \Psi(z_{1:t}) \end{aligned}$$

Therefore the VLB becomes:

$$\mathcal{L}(\theta, \phi, X) = \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})} \sum_{t=1}^T \log p_{\theta_x}(x_t|x_{1:t-1}, z_{1:t}) - \sum_{t=1}^T \mathbb{E}_{q_\phi(z_{1:t-1}|x_{1:T})} \left[ \mathbb{E}_{q_\phi(z_t|z_{1:t-1}, x_{1:T})} \log \frac{q_\phi(z_t|z_{1:t-1}, x_{1:T})}{p_{\theta_z}(z_t|z_{1:t-1}, x_{1:t-1})} \right] \quad (5.9)$$

$$= \sum_{t=1}^T \mathbb{E}_{q_\phi(z_{1:t}|x_{1:T})} \log p_{\theta_x}(x_t|x_{1:t-1}, z_{1:t}) - \sum_{t=1}^T \mathbb{E}_{q_\phi(z_{1:t-1}|x_{1:T})} \mathbb{KL}(q_\phi(z_t|z_{1:t-1}, x_{1:T}) || p_{\theta_z}(z_t|z_{1:t-1}, x_{1:t-1})) \quad (5.10)$$

As for the vanilla VAE, the VLB contains two terms.

- The first term is the reconstruction error. it is the sum over the time steps, of the average log likelihood the data at time  $t$ , given the approximate distribution of the past and present latent variables, and the past data.
- The second term is a regularization term, summing over the time steps the average divergence between the approximate posterior distribution of the latent variable at time  $t$ , and its real distribution.

As in vanilla VAE, the sampling over  $q_\phi$  requires the use of the "re parametrization trick" (see [11]), for  $\mathcal{L}(\theta, \phi, X)$  to be differentiable w.r.t.  $\theta, \phi$ .

Here is the summary regarding DVAE:

- **generative model**

$$p(x_{1:T}, z_{1:T}) = \prod_{t=1}^T p_{\theta_x}(x_t | x_{1:t-1}, z_{1:t}) p_{\theta_z}(z_t | z_{1:t-1}, x_{1:t-1}) \quad (5.11)$$

- **inference model**

$$q_{\phi}(z_{1:T} | x_{1:T}) = \prod_{t=1}^T q_{\phi}(z_t | z_{1:t-1}, x_{1:T}) \quad (5.12)$$

- **VLB for training**

$$\begin{aligned} \mathcal{L}(\theta, \phi, X) = & \sum_{t=1}^T \mathbb{E}_{q_{\phi}(z_{1:t} | x_{1:T})} \log p_{\theta_x}(x_t | x_{1:t-1}, z_{1:t}) \\ & - \sum_{t=1}^T \mathbb{E}_{q_{\phi}(z_{1:t-1} | x_{1:T})} \mathbb{KL} \left( q_{\phi}(z_t | z_{1:t-1}, x_{1:T}) || p_{\theta_z}(z_t | z_{1:t-1}, x_{1:t-1}) \right) \end{aligned} \quad (5.13)$$

## Deep Kalman Filter

The Kalman Filter is a well known model, widely used to denoise time series observations and make predictions. The latent variables form a Markov Chain, and all the probability distributions (ie encoder, decoder and transition model) are linear Gaussians. This allows to derive close form expressions for the solutions (Kalman filter and Kalman smoother).

In a **Deep Kalman Filter**, the temporal structure of the latent variables is still a Markov Chain. The probability models are still Gaussians, but with parameters mean and covariance learnt by neural networks.

More specifically, the DAG describing a Deep Kalman Filter is:

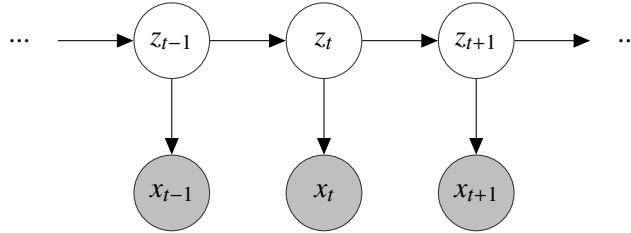


Figure 6.1: Probabilistic model of a Deep Kalman Filter

It is then particularly useful to use D-separation on the DAG to simplify the general DVAE expressions 5.11 and 5.12. Conditioning on  $z_t$  and  $z_{t-1}$  drives:

$$p_{\theta_x}(x_t|x_{1:t-1}, z_{1:t}) = p_{\theta_x}(x_t|z_t) \quad (6.1)$$

$$p_{\theta_z}(z_t|z_{1:t-1}, x_{1:t}) = p_{\theta_z}(z_t|z_{t-1}) \quad (6.2)$$

$$q_{\phi}(z_t|z_{1:t-1}, x_{1:T}) = q_{\phi}(z_t|z_{t-1}, x_{t:T}) \quad (6.3)$$

We then choose Gaussian distributions for  $p_{\theta_x}$ ,  $p_{\theta_z}$  and  $q_{\phi}$ , with mean and diagonal covariance, learnt by neural networks.

$$p_{\theta_x}(x_t|z_t) = \mathcal{N}(x_t|\mu_{\theta_x}(z_t), \text{diag } \sigma_{\theta_x}^2(z_t)) \quad (6.4)$$

$$p_{\theta_z}(z_t|z_{t-1}) = \mathcal{N}(z_t|\mu_{\theta_z}(z_{t-1}), \text{diag } \sigma_{\theta_z}^2(z_{t-1})) \quad (6.5)$$

$$q_{\phi}(z_t|z_{t-1}, x_{t:T}) = \mathcal{N}(z_t|\mu_{\phi}(z_{t-1}, x_{t:T}), \text{diag } \sigma_{\phi}^2(z_{t-1}, x_{t:T})) \quad (6.6)$$

Some other formulations of the approximate posterior (encoder) are possible. For example:

$$\begin{aligned} q_\phi(z_t|z_{t-1}, x_t) \\ q_\phi(z_t|z_{1:t}, x_{1:t}) \\ q_\phi(z_t|z_{1:T}, x_{1:T}) \end{aligned}$$

We have chosen 6.3 for the implementation, as it has the same formulation as the true posterior and respects the corresponding dependencies.

Taking note that:

$$q_\phi(z_{1:t}|x_{1:T}) = q_\phi(z_{1:t-1}|z_t, x_{1:T})q_\phi(z_t|x_{1:T})$$

And using D-Separation, the Evidence Lower Bound (ELBO) 5.13 simplifies into:

$$\mathcal{L}(\theta, \phi, X) = \sum_{t=1}^T \mathbb{E}_{q_\phi(z_{1:t}|x_{1:T})} \log p_{\theta_x}(x_t|z_t) - \sum_{t=1}^T \mathbb{E}_{q_\phi(z_{1:t-1}|x_{1:T})} \mathbb{KL}(q_\phi(z_t|z_{t-1}, x_{t:T}) || p_{\theta_z}(z_t|z_{t-1})) \quad (6.7)$$

$$= \sum_{t=1}^T \mathbb{E}_{q_\phi(z_t|x_{1:T})} \log p_{\theta_x}(x_t|z_t) - \sum_{t=1}^T \mathbb{E}_{q_\phi(z_{t-1}|x_{1:T})} \mathbb{KL}(q_\phi(z_t|z_{t-1}, x_{t:T}) || p_{\theta_z}(z_t|z_{t-1})) \quad (6.8)$$

As a summary:

#### Deep Kalman Filter

- **generative model**

$$p_{\theta_x}(x_t|z_t) = \mathcal{N}(x_t|\mu_{\theta_x}(z_t), \text{diag } \sigma_{\theta_x}^2(z_t)) \quad (6.9)$$

$$p_{\theta_z}(z_t|z_{t-1}) = \mathcal{N}(z_t|\mu_{\theta_z}(z_{t-1}), \text{diag } \sigma_{\theta_z}^2(z_{t-1})) \quad (6.10)$$

- **inference model**

$$q_\phi(z_t|z_{t-1}, x_{t:T}) = \mathcal{N}(z_t|\mu_\phi(z_{t-1}, x_{t:T}), \text{diag } \sigma_\phi^2(z_{t-1}, x_{t:T})) \quad (6.11)$$

- **VLB for training**

$$\mathcal{L}(\theta, \phi, X) = \sum_{t=1}^T \mathbb{E}_{q_\phi(z_t|x_{1:T})} \log p_{\theta_x}(x_t|z_t) - \sum_{t=1}^T \mathbb{E}_{q_\phi(z_{t-1}|x_{1:T})} \mathbb{KL}(q_\phi(z_t|z_{t-1}, x_{t:T}) || p_{\theta_z}(z_t|z_{t-1})) \quad (6.12)$$

The  $\mathbb{KL}(q_\phi || p_{\theta_z})$ 's have a close form, as the two distributions are Gaussians (see C)

From a code stand-point, following [14], we have used forward LSTM to encode sequences such as  $x_{1:t}$ , and backward LSTM to encode sequences such as  $x_{t:T}$ , as inputs into the MLP parametrizing the distributions.

For example:

$$\begin{aligned} \overleftarrow{g}_t &= \text{Backward LSTM}(\overleftarrow{g}_{t+1}, x_t) \text{ (encodes } x_{t:T}) \\ q_\phi(z_t|z_{t-1}, x_{t:T}) &= \mathcal{N}(z_t|\mu_\phi(z_{t-1}, \overleftarrow{g}_t), \text{diag } \sigma_\phi^2(z_{t-1}, \overleftarrow{g}_t)) \end{aligned}$$

The PyTorch implementation is described below:

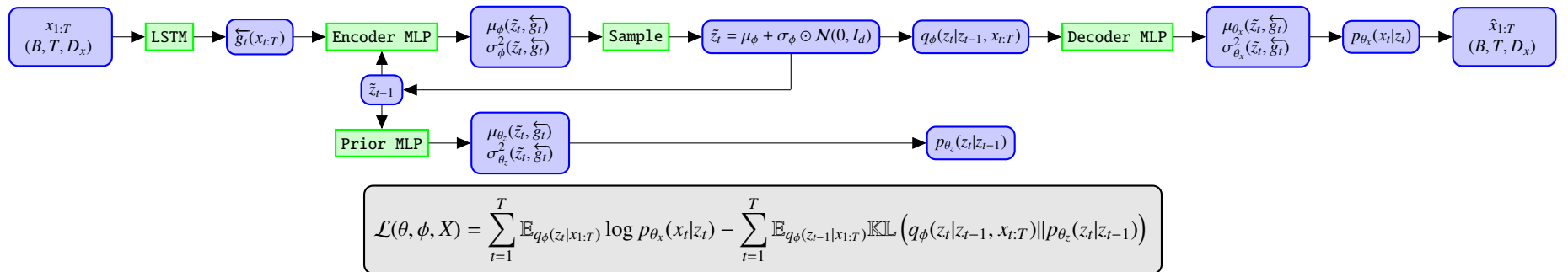


Figure 6.2: Deep Kalman Filter Model Architecture

## Variational Recurrent Neural Network

The VRNN is the most expressive DVAE, in that sense that the general expressions 5.11, 5.12 and VLB 5.13 can not be simplified.

The GPM of the VRNN assumes full connections between latent variables, and between observed variables, to account for the full unsimplified expressions. Specifically:

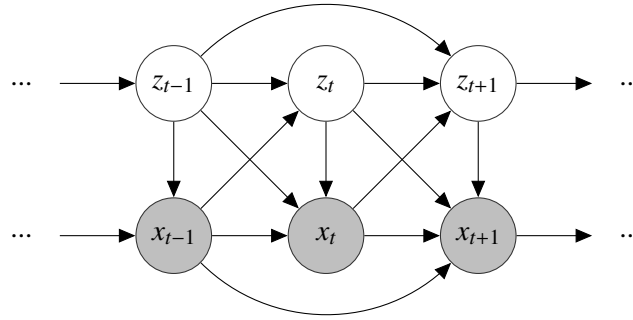


Figure 7.1: Probabilistic model of a Variational RNN

We remember that:

$$p(x_{1:T}, z_{1:T}) = \prod_{t=1}^T p_{\theta_x}(x_t | x_{1:t-1}, z_{1:t}) p_{\theta_z}(z_t | x_{1:t-1}, z_{1:t-1})$$

And posit Gaussian distributions with diagonal covariance and mean given by two networks:

$$p_{\theta_x}(x_t | x_{1:t-1}, z_{1:t}) = \mathcal{N}(x_t | \mu_{\theta_x}(x_{1:t-1}, z_{1:t}), \text{diag } \sigma_{\theta_x}^2(x_{1:t-1}, z_{1:t})) \quad (7.1)$$

$$p_{\theta_z}(z_t | z_{1:t-1}, x_{1:t-1}) = \mathcal{N}(z_t | \mu_{\theta_z}(z_{1:t-1}, x_{1:t-1}), \text{diag } \sigma_{\theta_z}^2(z_{1:t-1}, x_{1:t-1})) \quad (7.2)$$

The true posterior being

$$p(z_{1:T} | x_{1:T}) = \prod_{t=1}^T p(z_t | z_{1:t-1}, x_{1:T})$$

we choose the encoder with the same conditional dependencies and a Gaussian expression:

$$q_\phi(z_t|z_{1:t-1}, x_{1:T}) = \mathcal{N}(z_t|\mu_\phi(z_{1:t-1}, x_{1:T}), \text{diag } \sigma_\phi^2(z_{1:t-1}, x_{1:T}))$$

The VLB is:

$$\mathcal{L}(\theta, \phi, X) = \sum_{t=1}^T \left[ \mathbb{E}_{q_\phi(z_{1:t}|x_{1:T})} \log p_{\theta_x}(x_t|x_{1:t-1}, z_{1:t}) - \mathbb{E}_{q_\phi(z_{1:t-1}|x_{1:T})} \mathbb{KL}\left(q_\phi(z_t|z_{1:t-1}, x_{1:T}) || p_{\theta_z}(z_t|z_{1:t-1}, x_{1:t-1})\right) \right]$$

As a summary

#### Variational RNN

- **generative model**

$$p_{\theta_x}(x_t|x_{1:t-1}, z_{1:t}) = \mathcal{N}(x_t|\mu_{\theta_x}(x_{1:t-1}, z_{1:t}), \text{diag } \sigma_{\theta_x}^2(x_{1:t-1}, z_{1:t})) \quad (7.3)$$

$$p_{\theta_z}(z_t|z_{1:t-1}, x_{1:t-1}) = \mathcal{N}(z_t|\mu_{\theta_z}(z_{1:t-1}, x_{1:t-1}), \text{diag } \sigma_{\theta_z}^2(z_{1:t-1}, x_{1:t-1})) \quad (7.4)$$

- **inference model**

$$q_\phi(z_t|z_{1:t-1}, x_{1:T}) = \mathcal{N}(z_t|\mu_\phi(z_{1:t-1}, x_{1:T}), \text{diag } \sigma_\phi^2(z_{1:t-1}, x_{1:T})) \quad (7.5)$$

- **VLB for training**

$$\begin{aligned} \mathcal{L}(\theta, \phi, X) = & \sum_{t=1}^T \mathbb{E}_{q_\phi(z_{1:t}|x_{1:T})} \log p_{\theta_x}(x_t|x_{1:t-1}, z_{1:t}) \\ & - \sum_{t=1}^T \mathbb{E}_{q_\phi(z_{1:t-1}|x_{1:T})} \mathbb{KL}\left(q_\phi(z_t|z_{1:t-1}, x_{1:T}) || p_{\theta_z}(z_t|z_{1:t-1}, x_{1:t-1})\right) \end{aligned} \quad (7.6)$$

We have chosen a different implementation from [14] and used three different LSTM networks to encode  $z_{1:t}$ ,  $x_{1:t-1}$  and  $x_{t:T}$  respectively.

The PyTorch implementation is described below:



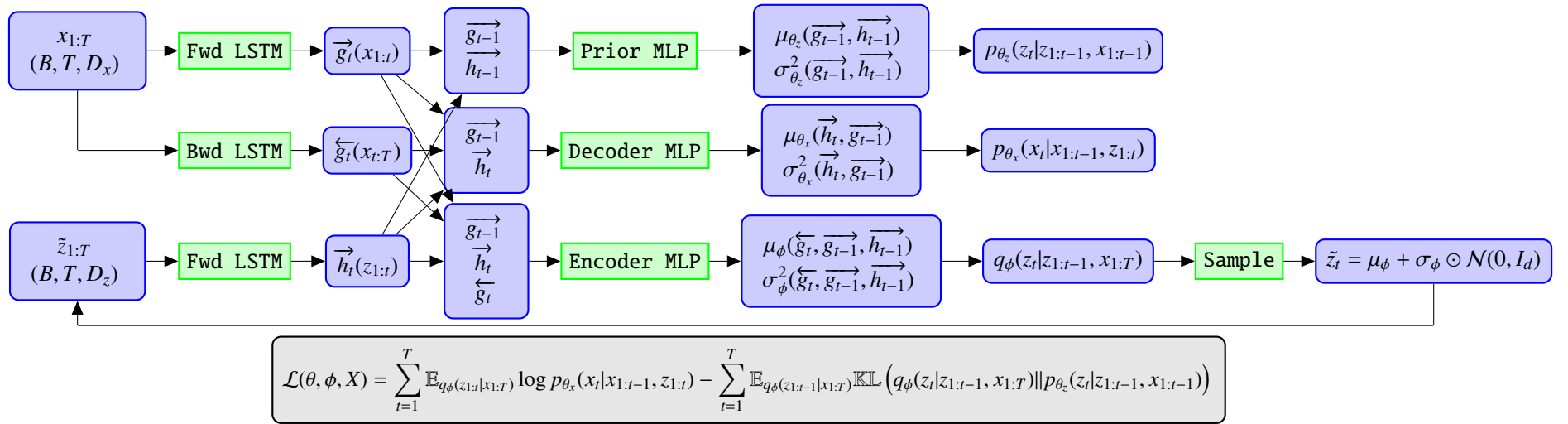


Figure 7.2: Variational RNN Model Architecture

## Gaussian Process Variational Auto Encoder

DVAEs are a natural and straightforward extension of VAEs to the time domain. However, the discretization of time comes with limitations. First, one has to choose a relevant time interval to sample the data, which can prove arbitrary if the observed process is not well known. Second, that time interval is fixed for training and inference, can not be changed depending on the time dynamics of the observed process, and can not account for different time scales.

It is therefore interesting to allow a continuous-time formulation of the prior of the latent variables, to provide additional flexibility and expressiveness. The natural and straightforward framework for such a continuous time prior is the GP, that constitutes the core of GP-VAEs. (A summary of GP can be found in [B](#)).

If the use of GPs for time-series modeling is not recent (see for example [\[6\]](#) and [\[7\]](#)), structuring a latent prior as a GP is somewhat newer. In [\[9\]](#), Casale and al. build a GP-VAE generative model to predict images with different objects and views. A specific kernel is designed for the task, taking advantage of the kernel construction rules and multiplying a view-based kernel by an object-based kernel. The kernel parameters are learnt with the inference model, and the covariance matrix of the kernel is built with a low-rank approximation ( $VV^T$ ) to reduce computation costs (naïvely in  $O(T^3)$ ). In [\[13\]](#), Fortuin and al. focus on time-series missing values imputations. A Cauchy kernel is used, which is an instance of a Rational Quadratic Kernel, that can be viewed as an infinite sum of Gaussian kernels over the space of lengthscales. The encoder  $q_\phi$  is a multivariate normal distribution, whose precision matrix is built multiplying a bi-band matrix and its transpose, again to reduce computation cost. [\[14\]](#) cites GP-VAE but remains focused on discrete-time models. The paper [\[16\]](#) establishes the Markovian nature of a GP as the solution of a linear SDE, which allows to significantly reduce the computation cost of the model.

The main insight is that the solution of a linear SDE is a Gaussian Process, as the transition probabilities given by the Fokker-Plank equations are Gaussian. Thus, GP-VAEs have the potential to express naturally many phenomena described by SDEs. We will review later those results, issued from stochastic calculus reference books such as [\[5\]](#), [\[12\]](#) and [\[18\]](#).

We now move to the GP-VAE model itself.

We can consider **data taken at irregular time intervals**. We change our notation accordingly, and note  $(t_1, \dots, t_{i-1}, t_i, t_{i+1}, \dots, t_T)$  the  $T$  times (or timestamps) considered.

The GPM of the GP-VAE is:

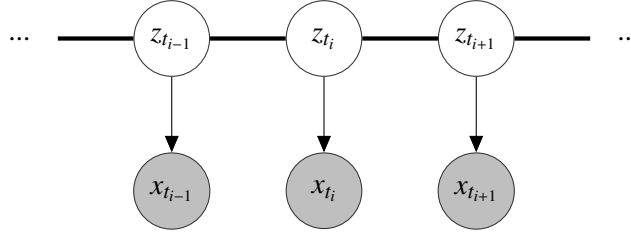


Figure 8.1: Probabilistic model of a GP-VAE

Where the thick black lines between latent variables define a fully connected graph : all latent variables are -a priori- correlated between each other in a Gaussian Process. (NB : this GPM is not, per say, a DAG in this regard. However, D-separation still applies for observed variables  $x_{t_i}$ ).

The joint distribution writes somehow differently from the one for DVAEs, as:

$$p(x_{t_1:t_T}, z_{t_1:t_T}) = p(z_{t_1:t_T})p(x_{t_1:t_T}|z_{t_1:t_T}) \quad (8.1)$$

$$= p(z_{t_1:t_T}) \prod_{i=1}^T p(x_{t_i}|x_{t_1:t_{i-1}}, z_{t_1:t_T}) \quad (8.2)$$

$$= p(z_{t_1:t_T}) \prod_{i=1}^T p(x_{t_i}|z_{t_i}) \quad (8.3)$$

The prior over the latent variables  $z_{t_i} \in \mathbb{R}^L$  is a set of scalar Gaussian Process over each of the dimension  $l \in \{1, \dots, L\}$  of the latent variables. Formally:

$$p_{\theta_z}(z_{t_1:t_T}^l) = \mathcal{GP}(m_{\theta_z,l}(t_1 : t_T), k_{\theta_z,l}(t_1 : t_T, t_1 : t_T)) \quad l = 1, \dots, L \quad (8.4)$$

where the  $m_{\theta_z,l}$  are the  $L$  mean functions of the GP priors (usually chosen constant null), and the  $k_{\theta_z,l}$  are the kernel functions of the GP priors.

We note at this point that:

- by design, each of the component of the  $z_{t_i}$  is a scalar GP, with correlation over time stamps. However, the different components of a  $z_{t_i}$  are not correlated between them. The correlation across dimensions is encoded into the observation model  $p_{\theta_x}(x_{t_i}|z_{t_i})$ .
- the kernels  $k_{\theta_z,l}$  can be chosen differently to account for different prior knowledge of the data sequence. In [13] for example, Fortuin and al. uses a set of Gaussian Kernels with different lengthscales.

Accordingly, the approximate posterior -encoder-  $q_\phi$  is a set of  $L$  Gaussian distributions of dimension  $T$ , each one accounting for a component of  $z_{t_i}$ . Formally :

$$q_\phi(z_{t_1:t_T}^l|x_{t_1:t_T}^l) = \mathcal{N}(m_\phi^l(x_{t_1:t_T}), \Sigma_\phi^l(x_{t_1:t_T})) \quad l = 1, \dots, L \quad (8.5)$$

$$= \mathcal{N}(m_\phi^l(x_{t_1:t_T}), \Lambda_\phi^l(x_{t_1:t_T})^{-1}) \quad (8.6)$$

$$= \mathcal{N}(m_\phi^l(x_{t_1:t_T}), L_\phi^l(x_{t_1:t_T})L_\phi^l(x_{t_1:t_T})^T) \quad (8.7)$$

where we have made explicit the different ways of defining the multivariate normal distribution, with its covariance matrix  $\Sigma_\phi^l$ , its precision matrix  $\Lambda_\phi^l$ , or with a Cholesky decomposition  $L_\phi^l$ .

The observation model, by D-separation, is:

$$p(x_{t_1:t_T}|z_{t_1:t_T}) = \prod_{i=1}^T p_{\theta_x}(x_{t_i}|z_{t_i}) \quad (8.8)$$

The log-likelihood of the data writes:

$$\log p(x_{t_1:t_T}) = \log \frac{p(x_{t_1:t_T}, z_{t_1:t_T})}{p(z_{t_1:t_T}|x_{t_1:t_T})} \quad (8.9)$$

As usual, we multiply by  $q_\phi(z_{t_1:t_T}|x_{t_1:t_T})$  and integrate over  $dz_{t_1:t_T}$  to form the VLB:

$$\log p(x_{t_1:t_T}) = \int q_\phi(z_{t_1:t_T}|x_{t_1:t_T}) \log \frac{p(x_{t_1:t_T}, z_{t_1:t_T})}{q_\phi(z_{t_1:t_T}|x_{t_1:t_T})} \frac{q_\phi(z_{t_1:t_T}|x_{t_1:t_T})}{p(z_{t_1:t_T}|x_{t_1:t_T})} dz_{t_1:t_T} \quad (8.10)$$

$$= \mathbb{E}_{q_\phi(z_{t_1:t_T}|x_{t_1:t_T})} \log \frac{p(x_{t_1:t_T}, z_{t_1:t_T})}{q_\phi(z_{t_1:t_T}|x_{t_1:t_T})} + \mathbb{KL}(q_\phi(z_{t_1:t_T}|x_{t_1:t_T})||p(z_{t_1:t_T}|x_{t_1:t_T})) \quad (8.11)$$

$$\geq \mathbb{E}_{q_\phi(z_{t_1:t_T}|x_{t_1:t_T})} \log \frac{p(x_{t_1:t_T}, z_{t_1:t_T})}{q_\phi(z_{t_1:t_T}|x_{t_1:t_T})} = \mathcal{L}(\theta, \phi, X) \quad (8.12)$$

Factoring in 8.1 and 8.8, we get:

$$\mathcal{L}(\theta, \phi, X) = \mathbb{E}_{q_\phi(z_{t_1:t_T}|x_{t_1:t_T})} \log \left[ \left( \prod_{i=1}^T p_{\theta_x}(x_{t_i}|z_{t_i}) \right) \frac{p_{\theta_z}(z_{t_1:t_T})}{q_\phi(z_{t_1:t_T}|x_{t_1:t_T})} \right] \quad (8.13)$$

$$= \sum_{i=1}^T \mathbb{E}_{q_\phi(z_{t_1:t_T}|x_{t_1:t_T})} \log p_{\theta_x}(x_{t_i}|z_{t_i}) - \mathbb{KL}(q_\phi(z_{t_1:t_T}|x_{t_1:t_T})||p_{\theta_z}(z_{t_1:t_T})) \quad (8.14)$$

We have  $\mathbb{E}_{q_\phi(z_{t_1:t_T}|x_{t_1:t_T})} f(z_{t_i}) = \mathbb{E}_{q_\phi(z_{t_i}|x_{t_i})} f(z_{t_i})$  for any  $f$ , so we get finally:

$$\mathcal{L}(\theta, \phi, X) = \sum_{i=1}^T \mathbb{E}_{q_\phi(z_{t_i}|x_{t_i})} \log p_{\theta_x}(x_{t_i}|z_{t_i}) - \mathbb{KL}(q_\phi(z_{t_1:t_T}|x_{t_1:t_T})||p_{\theta_z}(z_{t_1:t_T})) \quad (8.15)$$

We note that:

- the  $\mathbb{KL}$ -divergence is actually the sum of the  $L$   $\mathbb{KL}$ -divergences  $\mathbb{KL}(q_\phi^l||p_{\theta_z}^l)$ , which have a close form solution as both distributions are Gaussian. (see the well-known result C)
- the reconstruction loss term requires sampling from  $q_\phi(z_{t_i}|x_{t_i})$  using the reparameterization trick as usual.
- the GP priors  $p_{\theta_z}(z_{t_1:t_T})$  depend only on the time stamps  $t_1, \dots, t_T$ . If the kernel parameters are fixed -such as in [13]- then the priors can be computed before the training loop. If the kernel parameters are learnt with the weights of the neural nets (such as in [16]), then the computation must occur at each training iteration.

As a summary:

- **generative model**

$$p(x_{t_1:t_T}, z_{t_1:t_T}) = p(z_{t_1:t_T}) \prod_{i=1}^T p(x_{t_i} | z_{t_i}) \quad (8.16)$$

$$p_{\theta_z}(z_{t_1:t_T}^l) = \mathcal{GP}(m_{\theta_z, l}(t_1 : t_T), k_{\theta_z, l}(t_1 : t_T)) \quad l = 1, \dots, L \quad (8.17)$$

- **inference model**

$$q_{\phi}(z_{t_1:t_T}^l | x_{t_1:t_T}^l) = \mathcal{N}(m_{\phi}^l(x_{t_1:t_T}), \Sigma_{\phi}^l(x_{t_1:t_T})) \quad l = 1, \dots, L \quad (8.18)$$

$$= \mathcal{N}(m_{\phi}^l(x_{t_1:t_T}), \Lambda_{\phi}^l(x_{t_1:t_T})^{-1}) \quad (8.19)$$

$$= \mathcal{N}(m_{\phi}^l(x_{t_1:t_T}), L_{\phi}^l(x_{t_1:t_T}) L_{\phi}^l(x_{t_1:t_T})^T) \quad (8.20)$$

- **VLB for training**

$$\mathcal{L}(\theta, \phi, X) = \sum_{i=1}^T \mathbb{E}_{q_{\phi}(z_{t_i} | x_{t_1:t_T})} \log p_{\theta_x}(x_{t_i} | z_{t_i}) - \mathbb{KL}(q_{\phi}(z_{t_1:t_T} | x_{t_1:t_T}) || p_{\theta_z}(z_{t_1:t_T})) \quad (8.21)$$

The PyTorch implementation is described here:

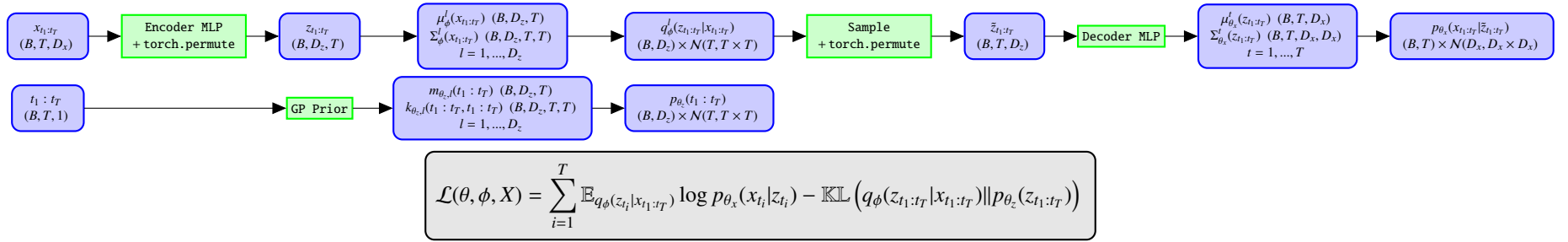


Figure 8.2: Gaussian Process VAE Model Architecture

## **Part III**

# **Notions on stochastic differential equations and their relationships to DVAEs**

This part intends to present a self-contained material on stochastic calculus. The goal is to highlight the relationships between SDEs and DVAEs. The part is designed so a complete beginner on the matter -with sufficient mathematical background though !- may grasp enough notions so the relationships appear natural. However, stochastic calculus is a full mathematical field in itself, and a detailed presentation is out of the scope of this report.

The interested reader will likely enjoy [5], [12], and refer to [18], for a fuller review of the subject.



## Stochastic processes basics

We posit the following in the rest of the doc, unless specified otherwise:

### Assumption 9.0.1: Main assumptions

1.  $(\Omega, \mathcal{F}, \mathbb{P})$  is a probabilistic space, with  $\Omega$  the universe,  $\mathcal{F}$  the tribe ( $\sigma$ -algebra) of events,  $\mathbb{P}$  a probability measure on  $(\Omega, \mathcal{F})$ .
2.  $(E, \mathcal{B}(E))$  is a measurable space, endowed with its Borelian  $\sigma$ -algebra.  $E$  will sometimes be referred to as the *state space*. It is typically the space where random variables will live, most of the time  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ .
3.  $T$  is the set of *times*, typically  $T = [0, a]$  with  $a > 0$ , or  $T = [0, +\infty[$ .

We view a stochastic process as a two-variable function from  $T \times \Omega$  to  $(E, \mathcal{B}(E))$ :

$$X : T \times \Omega \longrightarrow (E, \mathcal{B}(E)) \quad (9.1)$$

$$(t, \omega) \longmapsto X(t, \omega) \quad (9.2)$$

This leads to two complementary views:

**Stochastic process as a collection of random variables indexed by time** We view  $X$  as a collection of random variables  $(X_t)_{t \in T}$ :

$$\forall t \in T, X_t : \omega \mapsto X_t(\omega) \quad (9.3)$$

1. Each random variable  $X_t$  is defined on  $(\Omega, \mathcal{F}, \mathbb{P})$
2. In practice, only some events  $A \in \mathcal{F}$  can occur during the times  $[0, t] \in T$ . We will note  $\mathcal{F}_t$  the smallest  $\sigma$ -algebra that contains the set of events  $A$  that can occur during  $[0, t]$ .
3. As a result,  $X_t$  must be  $\mathcal{F}_t$ -measurable.

An alternate view arises when we fix  $\omega$  and allow  $t \in T$ :

**Stochastic process as a set of random trajectories** We view  $X$  as

$$X : \Omega \longrightarrow \mathbb{R}^T \quad (9.4)$$

$$\omega \longmapsto X(\omega) : t \mapsto X(t, \omega) \quad (9.5)$$

1.  $\forall \omega \in \Omega, X(\omega) \in \mathbb{R}^T$  is a random trajectory from  $T$  into  $\mathbb{R}$
2.  $\mathbb{R}^T$  should be endowed with an appropriate  $\sigma$ -algebra.

Those intuitions lead to a formal definition :

### Definition 9.0.2: Stochastic process

A stochastic process  $\mathbf{X}$  is defined as:

$$X = (\Omega, \mathcal{F}, (X_t)_{t \in T}, \mathbb{P}) \quad (9.6)$$

$$= (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in T}, (X_t)_{t \in T}, \mathbb{P}) \quad (9.7)$$

where:

- $\Omega$  is a set (universe of possibles).
- $\mathcal{F}$  is a  $\sigma$ -algebra of parts of  $\Omega$
- $\mathbb{P}$  is a probability measure on  $(\Omega, \mathcal{F})$
- $T \subset \mathbb{R}_+$  represents time
- $(\mathcal{F}_t)_{t \in T}$  is a **filtration**, ie an increasing family of sub- $\sigma$ -algebras of  $\mathcal{F}$  indexed by  $t : \forall 0 \leq s \leq t \in T, \mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$ .
- $(X_t)_{t \in T}$  is a family of RV defined on  $(\Omega, \mathcal{F})$  with values in a measurable space  $(E, \mathcal{E})$  or more simply  $(E, \mathcal{B}(E))$  (set  $E$  endowed with its Borelian  $\sigma$ -algebra).
- $(X_t)_{t \in T}$  is assumed **adapted to the filtration**  $(\mathcal{F}_t)_{t \in T}$ , meaning  $\forall t \in T, X_t$  is  $\mathcal{F}_t$ -measurable

Often, we will have  $(E, \mathcal{B}(E)) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ , and  $T = [0, a]$  or  $T = [0, +\infty[$  or  $T = \mathbb{N}$ .

Other important notions:

- **$\sigma$ -algebra generated by a random variable.** Let  $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (E, \mathcal{E})$  a random variable. The  $\sigma$ -algebra generated by  $X$  is the smallest  $\sigma$ -algebra that makes  $X$ -measurable. It is the set of "pre-images" of  $\mathcal{B}(E)$  by  $X$ . Formally;

$$\sigma(X) = \{A = X^{-1}(B), B \in \mathcal{E}\} \quad (9.8)$$

- **natural filtration of  $(X_t)_{t \in T}$ .** The set of  $\mathcal{F}_t = \sigma(X_s, s \leq t), \forall t \in T$ , ie the set of  $\sigma$ -algebras generated by the RV  $X_s$  for  $s \leq t$ , is called the natural filtration of  $(X_t)$

### Definition 9.0.3: Finite dimension laws of a stochastic process

Let  $X$  be a stochastic process,  $I = \{t_1, t_2, \dots, t_n\}, t_1 < t_2 < \dots < t_n$  a finite part of  $T$ , and  $X_I = (X_{t_1}, X_{t_2}, \dots, X_{t_n}) \in E^I$  a random vector.

Then, the **law of the random vector**  $X_I$  is the probability measure  $\mu_I$  image of  $\mathbb{P}$  by  $X_I : \Omega \rightarrow (E^I, \mathcal{B}(E^I))$

We also remind of Gaussian processes -with a view different from [6].

### Definition 9.0.4: Gaussian vector

Let  $\xi = (X_1, X_2, \dots, X_n)$  be a **centered** random vector in  $\mathbb{R}^n$  ( $\mathbb{E}(\xi) = 0$ ), verifying  $\xi \in L^2(\Omega, \mathcal{F}, \mathbb{P})$  (ie  $\forall i, \mathbb{E}(X_i^2) < \infty$ ). Let  $\Gamma = \mathbb{E}(X_i X_j)|_{i,j}$  its covariance matrix (definite positive).

$\xi$  is a **Gaussian vector** iff, equivalently:

1.  $\forall a_1, a_2, \dots, a_n \in \mathbb{R}^n$ ,  $\sum_{i=1}^n a_i X_i$  follows a normal law (ie is Gaussian)
2. or : the characteristic function of  $\xi$ ,  $\Phi_\xi(t) = \mathbb{E}(e^{i\langle \xi | t \rangle})$ , can be written

$$\Phi_\xi(t) = e^{-\frac{1}{2}\langle t | \Gamma t \rangle} \quad (t \in \mathbb{R}^n)$$

With notation  $\xi \sim \mathcal{N}(0, \Gamma)$ .

More generally,

$$Z \sim \mathcal{N}(m, \Gamma) \iff \Phi_Z(t) = e^{i\langle m | t \rangle} e^{-\frac{1}{2}\langle t | \Gamma t \rangle}$$

### Definition 9.0.5: Gaussian Process

Let  $X$  be a stochastic process taking its values in  $E = \mathbb{R}^n$ .

$X$  is said to be a **Gaussian process** iff all its finite dimension laws are Gaussian.

If  $X = (X_t)_{t \in T}$  is a real-valued Gaussian process ( $E = \mathbb{R}$ ), we define  $\forall s, t \in T$ :

- $m(t) = \mathbb{E}(X_t)$  is the **mean of the Gaussian process**
- $\Gamma(t, s) = \mathbb{E}((X_t - m(t))(X_s - m(s)))$  is the **covariance of the Gaussian process**

Reciprocally,

### Theorem 9.0.6: Existence of a Gaussian process

Let  $m : T \rightarrow \mathbb{R}$ , and  $\Gamma : T \times T \rightarrow \mathbb{R}$  be two functions such that:

$\forall I = \{t_1, t_2, \dots, t_n\}$  finite part of  $T$   $\Gamma_I = \Gamma(t_i, t_j)|_{i,j}$  is **symetric definite positive**

**then** there exists a Gaussian process  $X = (X_t)_{t \in T}$ , that is unique at a near equivalence, such that:

$$\begin{aligned} \forall I = \{t_1, t_2, \dots, t_n\} \subset T, X_I = \{X_{t_1}, X_{t_2}, \dots, X_{t_n}\} &\sim \mathcal{N}(m_I, \Gamma_I), \\ m_I &= (m(t_1), m(t_2), \dots, m(t_n)) \end{aligned}$$

*Proof for Theorem.*

■ p19

We will not cover notion such as stopping times or martingales, even they are central in stochastic calculus. Please refer to [5] for a detailed presentation.

The foundational stochastic process for stochastic calculus is the **brownian motion**:

### Definition 9.0.7: Brownian motion

A real-valued stochastic process  $B = (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, (B_t)_{t \geq 0}, \mathbb{P})$  is called **Brownian motion** iff:

- $B_0 = 0$   $\mathbb{P}$ -a.s.
- $\forall 0 \leq s \leq t$ , the random variable  $B_t - B_s$  is independent from  $\mathcal{F}_s$ .
- $\forall 0 \leq s \leq t$ ,  $B_t - B_s \sim \mathcal{N}(0, t - s)$

Meaning : the process  $B$  starts from 0, its increments are independent from the past, and follow a centered normal law of variance equal to the length of the time interval.

When  $(\mathcal{F}_t)_{t \geq 0}$  is the natural filtration of  $(B_t)_{t \geq 0}$ ,  $B$  is said to be a **natural Brownian motion**

A fundamental property of the Brownian motion is the following:

### Theorem 9.0.8: Gaussian characterization of the Brownian motion

1. Let  $B = (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, (B_t)_{t \geq 0}, \mathbb{P})$  be a Brownian motion. Then  $B$  verifies:

- $B_0 = 0$   $\mathbb{P}$ -a.s.
- $\forall 0 \leq t_1 < t_2 < \dots < t_n$ ,  $(B_{t_1}, B_{t_2}, \dots, B_{t_n})$  is a centered Gaussian vector.
- $\forall s, t \geq 0$ ,  $\mathbb{E}(B_s B_t) = \min(s, t)$

This means that  $B$  is a real centered Gaussian process, of covariance function  $\Gamma(t, s) = \min(s, t)$

2. Conversely, if  $B$  verifies the three properties above, then  $(\Omega, \mathcal{F}, (\tilde{\mathcal{F}}_t)_{t \geq 0}, (B_t)_{t \geq 0}, \mathbb{P})$  is a natural Brownian motion (with  $(\tilde{\mathcal{F}}_t)_{t \geq 0}$  the natural filtration of the family  $(B_t)_{t \geq 0}$ ).

*Proof for Theorem.*

see [5]

A second fundamental property of the Brownian motion is that its **quadratic variation** is non-zero. More formally,

### Definition 9.0.9: Variations of a function

Let  $f : [a, b] \rightarrow \mathbb{R}$  a function (with  $a, b \in \mathbb{R}$ ). Let  $\pi = \{a = t_0 < t_1 < \dots < t_n = b\}$  a subdivision of  $[a, b]$ .

1. The **variation of  $f$  along  $\pi$**  is  $V_\pi = \sum_{k=0}^{n-1} |f(t_{k+1}) - f(t_k)|$
2. The **total variation of  $f$  on  $[a, b]$**  is  $V_{[a,b]}^\pi = \sup_\pi V_\pi$ .  $f$  is said to have a **bounded variation** if  $V_{[a,b]}^\pi < \infty$ .
3. The **quadratic variation of  $f$  along  $\pi$**  is  $V_\pi^{(2)} = \sum_{k=0}^{n-1} |f(t_{k+1}) - f(t_k)|^2$
4. The **total quadratic variation of  $f$  along  $[a, b]$**  is :

$$[f]_{a,b} = \lim_{|\pi| \rightarrow 0} V_\pi^{(2)}$$
$$\text{with } |\pi| = \max_k |t_{k+1} - t_k|$$

The total quadratic variation has a slightly different definition from the total variation.

### Theorem 9.0.10: Quadratic variation of a Brownian motion

Let  $B = (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, (B_t)_{t \geq 0}, \mathbb{P})$  a Brownian motion. Then  $\forall 0 < s \leq t$

$$\lim_{|\pi| \rightarrow 0} V_\pi^{(2)} = t - s \text{ in } L^2 \quad (9.9)$$

$$\lim_{|\pi| \rightarrow 0} \mathbb{E}(|V_\pi^{(2)} - (t - s)|^2) = 0 \quad (9.10)$$

*Proof for Theorem.*

see [5]

The reader will refer to [5] for proofs of existence, continuity, and nowhere-differentiability of the Brownian motion.

Last, we introduce the notion of **Markov Process**, which is a generalization of the Markov Chains to continuous time.

Intuitively, building on the discrete-time Markov chain, a **Markov Process** is a stochastic process in which the behavior at time  $t$  given the information available up to time  $s$  (with  $0 < s < t$ ) depends only on the most recent past, ie the information at time  $s$  only.

In other words, for  $s < t, \in T$ , the law of  $X_t | \mathcal{F}_s$  depends only on  $X_s$ .

The intuition is then:

### Assumption 9.0.11: Intuition of a Markov Process

The stochastic process  $X$  is called a **Markov Process** if:

$$\forall s, t \in T, s < t, \forall A \in \mathcal{B}_E, \mathbb{P}(X_t \in A | \mathcal{F}_s) = \mathbb{P}(X_t \in A | X_s)$$

We note **the transition probability of, given a start from  $x$  at time  $s$ , to reach  $A \in \mathcal{B}_E$  at time  $t$** :

$$\mathbb{P}(X_t \in A | X_s = x) = P_{s,t}(x, A) \quad (9.11)$$

We see that  $A \mapsto P_{s,t}(x, A)$  is a **probability measure** on  $\mathcal{B}_E$ , that we note:

$$P_{s,t}(x, dy) : \mathcal{B}_E \rightarrow [0, 1] \quad (9.12)$$

$$A \mapsto P_{s,t}(x, A) \quad (9.13)$$

We now consider the space  $C = \{f : E \rightarrow \mathbb{R}, \text{borelian, bounded}\}$ , and the operator  $P_{s,t}$  from  $C$  into  $C$  defined by:

$$P_{s,t} : C \rightarrow C \quad (9.14)$$

$$f \mapsto P_{s,t}f : x \mapsto P_{s,t}f(x) = \int_E P_{s,t}(x, dy) f(y) \quad (9.15)$$

$$P_{s,t}f(x) = \int_E f(y) P_{s,t}(x, dy) = \mathbb{E}(f(X_t) | X_s = x) \quad (9.16)$$

Taking  $f = \mathbb{1}_A$ , we recover 9.11 from 9.16.

We are now equipped to define:

**Definition 9.0.12: Transition kernels**

A family  $(P_{s,t})_{s < t, t \in T}$  of applications  $(E, \mathcal{B}_E) \rightarrow [0, 1]$  is said to be a family of **transition kernels** iff:

- $\forall s < t, \forall A \in \mathcal{B}_E, P_{s,t}(\bullet, A) : x \mapsto P_{s,t}(x, A)$  is measurable.
- $\forall s < t, \forall x \in E, P_{s,t}(x, \bullet) : A \mapsto P_{s,t}(x, A)$  is a probability measure on  $\mathcal{B}_E$ .
- The **Chapman Kolmogorov** property holds:

$$\forall x \in E, \forall A \in \mathcal{B}_E, \forall s < t < u, P_{s,u}(x, A) = \int_E P_{s,t}(x, dy) P_{t,u}(y, A) \quad (9.17)$$

That is : we start from  $x$  at time  $s$ , we arrive at a random  $y$  (with some probability distribution over  $y$ ) at some intermediate time  $t$ , and then we start from  $y$  to reach  $A$  at time  $u$ .

Considering  $P_{s,t}$  as operators (9.16), then Chapman-Kolmogorov writes:

$$P_{s,u} = P_{s,t} P_{t,u} \quad (9.18)$$

And define a **Markov Process**:

**Definition 9.0.13: Markov Process**

A stochastic process  $X$  is said to be a Markov Process with transition kernels  $\{P_{s,t}; s, t \in T, s < t\}$  iff  $\forall f : E \rightarrow \mathbb{R}$  Borelian and bounded, and  $\forall s < t \in T$  we have:

$$\mathbb{E}(f(X_t) | \mathcal{F}_s) = P_{s,t} f(X_s) \quad \mathbb{P} - \text{a.s} \quad (9.19)$$

The transition kernels  $P_{s,t}$  are also called transition probabilities.

The law of  $X_0$  is a probability measure  $\nu$  over  $\mathcal{B}_E$  defined by:

$$\nu(A) = \mathbb{P}(X_0 \in A) \quad (9.20)$$

and called **initial law of the process**.

Again, if we take  $f = \mathbb{1}_A$ , we recover  $\mathbb{P}(X_t \in A | \mathcal{F}_s) = \mathbb{P}(X_t \in A | X_s)$ .

**Definition 9.0.14: Homogeneous Markov Process**

A Markov Process  $X$  is said to be **homogeneous** if its transition kernels family  $P_{s,t}$  depends only on  $t - s$ . ie:

$$\forall s < t, P_{s,t} = P_{0,t-s} := P_u \quad (u = t - s). \quad (9.21)$$

We can now compute the finite dimension laws of a Markov Process:

### Theorem 9.0.15: Finite dimension laws of a Markov Process

Let  $X$  be a Markov Process of initial law  $\nu$  and probability transitions  $P_{s,t}$ . For all finite sequence of times  $0 = t_0 < t_1 < \dots < t_k$ , and for all set of borelian bounded functions  $f_i : E \rightarrow \mathbb{R}, 0 \leq i \leq k$ , we have:

$$\mathbb{E}(f_0(X_0)f_1(X_1)\dots f_k(X_{t_k})) = \quad (9.22)$$

$$\int_E \nu(dx_0)f(x_0) \int_E P_{0,t_1}(x_0, dx_1)f_1(x_1)\dots \int_E P_{t_{k-1},t_k}(x_{k-1}, dx_k)f_k(x_k) \quad (9.23)$$

And  $\forall A_0, A_1, \dots, A_k \in \mathcal{B}_E$ , with  $f_i = \mathbb{1}_{A_i}$ :

$$\mathbb{P}(X_0 \in A_0, X_1 \in A_1, \dots, X_k \in A_k) = \int_{A_0} \nu(dx_0) \int_{A_1} P_{0,t_1}(x_0, dx_1)\dots \int_{A_k} P_{t_{k-1},t_k}(x_{k-1}, dx_k) \quad (9.24)$$

**Proof for Theorem.**

see [5] p92

Let  $(C_o(E), \|\cdot\|_\infty)$  be the Banach space<sup>1</sup> of functions  $f : E \rightarrow \mathbb{R}$  continuous, s.t.  $f \xrightarrow{\infty} 0$ .

Let  $(P_t)_{t \geq 0}$  be a family of positive operators<sup>2</sup>.

### Definition 9.0.16: Feller semi-group

$(P_t)_{t \geq 0}$  is said to be a **Feller semi-group** if:

- $P_0 = I_d$  and  $\forall t \geq 0, \|P_t\| \leq 1$
- $\forall t, t' \geq 0, P_t P_{t'} = P_{t+t'}$
- $\forall f \in C_o(E), \lim_{t \downarrow 0} \|P_t f - f\|_\infty = 0$

An homogeneous Markov process on  $E$  is said process of Feller if its semi-group is of Feller.

### Definition 9.0.17: Infinitesimal Generator

$X$  a process of Feller. Let  $f \in C_o(E)$  be such that the limit below exists in  $C_o(E)$ :

$$\lim_{t \downarrow 0} \frac{1}{t} (P_t f - f) = A f \quad (9.25)$$

then  $f$  is said to be in the domain  $D_A$  of  $A$  operator defined by 9.25.

**$A$  is called infinitesimal generator of the semi-group  $(P_t)_{t \geq 0}$ .**

Then we can write:

$$\mathbb{E}(f(X_{s+h})|\mathcal{F}_s) = f(X_s) + h A f(X_s) + o(h) \quad (9.26)$$

where  $o(h)$  depends only on  $f$ .

Regarding the Brownian motion, we have:

<sup>1</sup>Banach space : complete normed vector space

<sup>2</sup> $f \geq 0 \implies P_t f \geq 0$

### Theorem 9.0.18: Semi-group of the Brownian motion

Let  $B = (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in T}, (B_t)_{t \in T}, \mathbb{P})$  be a Brownian motion on  $\mathbb{R}$ . Then **B is an homogeneous Markov Process on  $\mathbb{R}$** , of initial law  $\nu = \delta_0$ , and whose semi-group is given by (for any  $f : E \rightarrow \mathbb{R}$  Borelian bounded):

$$P_t f(x) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{1}{2} \frac{(x-y)^2}{t}\right) f(y) dy \quad (9.27)$$

Or equivalently:

$$\forall A \in \mathcal{B}_{\mathbb{R}}, \quad P_t(x, A) = \int_A \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{1}{2} \frac{(x-y)^2}{t}\right) dy \quad (9.28)$$

That is,  $P_t(x, dy)$  is the Gaussian measure centered in  $x$ , of variance  $t$ .

***Proof for Theorem.***

see [5] p93

■



The idea is to consider stochastic processes  $Y_t$  whose "infinitesimal increments"  $dY_t$  (for  $t \in [a, b]$ ) are of the form  $dY_t = X_t dB_t$ , where  $dB_t$  is the infinitesimal increment of the Brownian motion  $B$ , and  $X = (X_t)_{t \in [a, b]}$  is a process adapted to the filtration  $(\mathcal{F}_t)_{t \in [a, b]}$  and "smooth enough". The infinitesimal Brownian increment  $dB_t$  has a non-null quadratic variation, which will lead eventually to the Itô's formula. The stochastic integral is defined on elementary stochastic processes, then extended to broader classes of stochastic processes.

### Definition 10.0.1: Elementary process

A stochastic process  $X = (X_s)_{s \in [a, b]}$  is called **elementary** if there exists a subdivision  $a = t_0 < t_1 < \dots < t_n = b$  of  $[a, b]$ , such that:

$$\forall t \in [a, b], \forall \omega \in \Omega, X_t(\omega) = \sum_{i=0}^{n-1} X_i(\omega) \mathbf{1}_{[t_i, t_{i+1}[}(t)$$

with  $\forall i \in \{0, 1, \dots, n-1\}$ ,  $X_i$  is  $\mathcal{F}_{t_i}$ -measurable.

This means that, in each interval  $[t_i, t_{i+1}[$ ,  $X_t(\omega)$  is independent of  $t$  and  $X_t(\omega) = X_i(\omega)$ .

We define  $\mathcal{E}$  (resp.  $\mathcal{E}_n, n > 0$ ) the set of all elementary processes on  $[a, b]$  (resp. the subset of the  $X \in \mathcal{E}$ ) such that all  $X_i$  have a finite moment  $\mathbb{E}X_i < \infty$  (resp.  $\mathbb{E}(|X_i|^n) < \infty$ )

### Definition 10.0.2: Stochastic integral of an elementary process

Let  $X \in \mathcal{E}$ , ie

$$X_t(\omega) = \sum_{i=0}^{n-1} X_i(\omega) \mathbf{1}_{[t_i, t_{i+1}[}(t)$$

The stochastic integral of  $X$  is the real random variable :

$$\int_a^b X_t dB_t := \sum_{i=0}^{n-1} X_i(B_{t_{i+1}} - B_{t_i})$$

### Proposition 10.0.3

1. **linearity**

$$\forall X, Y \in \mathcal{E}, \forall \lambda, \mu \in \mathbb{R}, \int_a^b (\lambda X_t + \mu Y_t) dB_t = \lambda \int_a^b X_t dB_t + \mu \int_a^b Y_t dB_t$$

2. **centering.** If  $X \in \mathcal{E}_1$ , then  $\int_a^b X_t dB_t \in L^1(\Omega, \mathcal{F}, \mathbb{P})$  and

$$\mathbb{E} \left( \int_a^b X_t dB_t \right) = 0$$

3. **membership in  $L^2(\Omega, \mathcal{F}, \mathbb{P})$ .** if  $X \in \mathcal{E}_2$ , then  $\int_a^b X_t dB_t \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ , and

$$\mathbb{E} \left[ \left( \int_a^b X_t dB_t \right)^2 \right] = \mathbb{E} \left( \int_a^b X_t^2 dt \right)$$

4. **corollary.** The application  $I$  is an isometry:

$$I : \mathcal{E}_2 \subset L^2([a, b] \times \Omega, \mathcal{B}_{[a,b]} \otimes \mathcal{F}, dt \otimes d\mathbb{P}) \rightarrow L^2(\Omega, \mathcal{F}, \mathbb{P})$$

$$X \mapsto I(X) := \int_a^b X_t dB_t$$

Let  $B = (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, (B_t)_{t \geq 0}, \mathbb{P})$  be a continuous Brownian motion.

Let  $X = (X_t)_{t \in [a,b]}$  a stochastic process defined on  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ , filtered space of  $B$ , restricted to  $[a, b]$ .

We define spaces of "smooth" stochastic processes  $X$  which allow to generalize the construction of the integral processes.

NB : in all the following,  $\int_a^b X_t^2 dt$  is the random variable defined by  $\forall \omega \in \Omega, \left( \int_a^b X_t^2 dt \right)(\omega) = \int_a^b X_t^2(\omega) dt$ .

### Definition 10.0.4: Space $M^2$

$X \in M^2$  if:

- $X$  is progressively measurable <sup>a</sup>
- And:

$$\mathbb{E} \left( \int_a^b X_t^2 dt \right) < +\infty$$

$M^2$  is a Hilbert space, with norm:

$$\|X\|_{M^2} := \|X\|_{L^2([a,b] \times \Omega, dt \otimes d\mathbb{P})} \quad (10.1)$$

$$\mathbb{E} \left( \int_a^b X_t^2 dt \right) = \int_{[a,b] \times \Omega} X_t(\omega)^2 dt d\mathbb{P} \quad (10.2)$$

<sup>a</sup>ie  $X$  is said to be progressively measurable w.r.t. a filtration  $(\mathcal{F}_t)_{t \in T}$  if  $\forall s \in T$ , the application  $(t, \omega) \mapsto X_t(\omega)$  is measurable from  $([0, s] \times \Omega, \mathcal{B}_{[0,s]} \times \mathcal{F}_s)$  to  $(E, \mathcal{B}_E)$ .

**Definition 10.0.5: Space  $\Lambda^2$** 

$X \in \Lambda^2$  if:

- $X$  is progressively measurable
- and:

$$\int_a^b X_t^2 dt < \infty, \quad \mathbb{P} - a.s.$$

**Proposition 10.0.6**

$$\mathcal{E}_2 \subset M^2 \subset \Lambda^2$$

**Theorem 10.0.7: Density of  $\mathcal{E}_2$  in  $M^2$** 

$\mathcal{E}_2$ , space of the square-integrable elementary processes, is a dense subspace of  $M^2$  :

$$\forall X \in M^2, \exists (X^{(n)})_{n \in \mathbb{N}} \in \mathcal{E}_2, \text{ s.t. } \lim_{n \rightarrow +\infty} \|X^{(n)} - X\|_{M^2} = 0$$

*Proof for Theorem.*

p122

**Theorem 10.0.8: Extension of the stochastic integral to  $M^2$** 

For  $X \in M^2$ , we define  $I(X) = \int_a^b X_t dB_t$  by:

$$\text{Let } (X^{(n)})_{n \in \mathbb{N}} \in \mathcal{E}_2, \text{ s.t. } \lim_{n \rightarrow +\infty} \|X^{(n)} - X\|_{M^2} = 0$$

$$\lim_{n \rightarrow +\infty} \int_a^b X_t^{(n)} dB_t \stackrel{L^2}{=} \int_a^b X_t dB_t$$

*Proof for Theorem.*

p123

This means that, to determine  $\int X_t dB_t$  with  $X \in M^2$ , we need to find a sequence  $X_t^{(n)} \in \mathcal{E}_2$  that converges (in  $L_2$ ) towards  $X$ .

### Proposition 10.0.9

Properties carry from  $\mathcal{E}_n$  to  $M^2$ , most notably:

1. **centering.** If  $X \in M^2$ , then  $\int_a^b X_t dB_t \in L^1(\Omega, \mathcal{F}, \mathbb{P})$  and

$$\mathbb{E} \left( \int_a^b X_t dB_t \right) = 0$$

2. **membership in  $L^2(\Omega, \mathcal{F}, \mathbb{P})$ .** if  $X \in M^2$ , then  $\int_a^b X_t dB_t \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ , and

$$\mathbb{E} \left( \left( \int_a^b X_t dB_t \right)^2 \right) = \mathbb{E} \left( \int_a^b X_t^2 dt \right)$$

3. **covariance - dot product conservation by isometry  $I$ .**  $\forall X, Y \in M^2$ ,

$$\mathbb{E} \left[ \left( \int_a^b X_t dB_t \right) \left( \int_a^b Y_t dB_t \right) \right] = \mathbb{E} \left( \int_a^b X_t Y_t dt \right)$$

The extension of the stochastic integral to processes in  $\Lambda_2$  also uses the convergence (this time, in probability) of a sequence  $X_t^{(n)} \in \mathcal{E}$  towards  $X$ . Then the integrals  $\int_a^b X_t^{(n)} dB_t$  converge towards  $\int_a^b X_t dB_t$  also in probability. More formally:

### Proposition 10.0.10

$\forall X \in \Lambda^2, \exists (X^{(n)})_{n \in \mathbb{N}} \in \mathcal{E}$  s.t.:

$$\lim_{n \rightarrow \infty} \int_a^b (X_t - X_t^{(n)})^2 dt = 0, \mathbb{P} - a.s.$$

So  $X_t^{(n)} \xrightarrow{\text{probability}} X_t$  also.

### Theorem 10.0.11: Extension of the stochastic integral to $\Lambda^2$

For  $X \in \Lambda^2$ , with  $(X^{(n)})_{n \in \mathbb{N}} \in \mathcal{E}$  st  $\lim_{n \rightarrow \infty} \int_a^b (X_t - X_t^{(n)})^2 dt = 0, \mathbb{P} - a.s.$ , then the sequence of random variables  $\int_a^b X_t^{(n)} dB_t$  converges in probability towards a random variable that is independent of  $(X^{(n)})$ :

$$\int_a^b X_t^{(n)} dB_t \xrightarrow{\text{probability}} I(X) := \int_a^b X_t dB_t$$

And, if  $X \in M^2$ , this definition is coincident with the definition of the stochastic integral in  $M^2$ .

**Proof for Theorem.**

NB : recall that the convergence in probability is defined by:

$$Z_n \xrightarrow[\text{probability}]{} Z \iff \forall \epsilon > 0, \lim_{n \rightarrow \infty} \mathbb{P}(|Z_n - Z| > \epsilon) = 0$$

### Theorem 10.0.12: Sums of Riemann-Stieltjes

Let  $X \in \Lambda^2$  be a **continuous** process. Then for all sequence of subdivisions  $(\pi_n)_{n \in \mathbb{N}}$ ,  $\pi_n = \{a = t_{n,0} < t_{n,1} < \dots < t_{n,m_n} = b\}$ , that verifies  $|\pi_n| \xrightarrow{n \rightarrow \infty} 0$  (the "step" of the subdivisions converges to 0), then:

$$\sum_{i=0}^{m_n-1} X_{t_{n,i}}(B_{t_{n,i+1}} - B_{t_{n,i}}) \xrightarrow{n \rightarrow \infty} \int_a^b X_t dB_t \quad \text{in probability}$$

*Proof for Theorem.*

p129

**Assumption 11.0.1: context**

In all the following,

- let  $B = (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, (B_t)_{t \geq 0}, \mathbb{P})$  be a continuous Brownian motion
- let  $T > 0$  a fixed time
- $\Lambda^p(0, T)$ ,  $p \geq 1$  is the set of progressively measurable processes  $X$  that verify  $\{t \mapsto X_t(\omega)\} \in L^p([0, T])$   $\mathbb{P}$ -a.s.
- $M^2([0, T])$  the set of progressively measurable processes  $X$  such that  $\mathbb{E} \left( \int_0^T X_t^2 dt \right) < \infty$   $\mathbb{P}$ -a.s.
- we will always consider the continuous version of the stochastic integrals

**Definition 11.0.2: Itô's process**

A stochastic process  $X = (X_t)_{t \geq 0}$  defined on  $(\Omega, \mathcal{F}, \mathbb{P})$  and adapted to the filtration  $(\mathcal{F}_t)_{t \in [0, T]}$  is called **Itô's process** if it exists two stochastic processes  $a_s \in \Lambda^1([0, T])$  and  $b_s \in \Lambda^2([0, T])$  such that:

$$\forall t \in [0, T], X_t = X_0 + \int_0^t a_s ds + \int_0^t b_s dB_s \quad (11.1)$$

Then we say that  $X$  **admits the stochastic differential** :

$$dX_t = a_t dt + b_t dB_t \quad (11.2)$$

The Itô's processes are stable by linear combination and multiplication - the Itô's processes set has an algebra structure.

### Theorem 11.0.3: Stochastic differential of a product, integration by parts

Let  $X, Y$  be two Itô's processes on  $[0, T]$ :

$$\begin{aligned}dX_t &= a_t^{(1)} dt + b_t^{(1)} dB_t \\dY_t &= a_t^{(2)} dt + b_t^{(2)} dB_t\end{aligned}$$

Then  $XY = (X_t Y_t)_{t \in [0, T]}$  is also a Itô's process and:

$$d(X_t Y_t) = X_t dY_t + Y_t dX_t + b_t^{(1)} b_t^{(2)} dt \quad (11.3)$$

The last term is the Itô's term.

*Proof for Theorem.*

book p155 for a clean proof, next chapter for a heuristic proof

### Theorem 11.0.4: Itô's formula

An Itô's process remains an Itô's process when it is transformed by a deterministic function that is "smooth enough".

Let  $X$  be a Itô's process on  $[0, T]$  :  $dX_t = a_t dt + b_t dB_t$ .

Let:

$$\begin{aligned}f &: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \\(x, t) &\mapsto f(x, t)\end{aligned}$$

be  $C^{2,1} : C^2$  in  $x$ , and  $C^1$  in  $t$ .

Then  $(f(X_t, t))_{t \in [0, T]}$  is also an Itô's process and:

$$d(f(X_t, t)) = \frac{\partial f}{\partial t}(X_t, t) dt + \frac{\partial f}{\partial x}(X_t, t) dX_t + \frac{1}{2} \frac{\partial^2 f}{\partial x^2}(X_t, t) b_t^2 dt \quad (11.4)$$

The last term is Itô's complementary term.

*Proof for Theorem.*

see book p159 for a clean proof, and next chapter for a heuristic proof

## 11.1 Stochastic differential equations - SDE

Let  $0 \leq a < b \in \mathbb{R}$ .

### Definition 11.1.1: Stochastic differential equation

We call **stochastic differential equation (SDE)** on  $[a, b]$ , with the initial data  $\xi_a$ , a relation such as:

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t \quad (11.5)$$

$$X_a = \xi_a \quad (11.6)$$

where:

- $X$  is an Itô's process on  $[a, b]$  (the unknown)
- $\xi_a$  is a given random variable,  $\mathcal{F}_a$ -measurable
- $\mu(x, t)$  and  $\sigma(x, t)$  are two given functions defined over  $\mathbb{R} \times [a, b] \rightarrow \mathbb{R}$

Solving the SDE is finding  $X$  Itô's process on  $[a, b]$  such that:

$$X_t = \xi_a + \int_a^t \mu(X_s, s)ds + \int_a^t \sigma(X_s, s)dB_s \quad \forall t \in [a, b]$$

NB : a SDE may not have a solution.

### Theorem 11.1.2: Existence and Unicity of a solution of a SDE

Let's assume

- $\sigma : t \mapsto \sigma(0, t)$  and  $\mu : t \mapsto \mu(0, t)$  bounded on  $[a, b]$
- $\sigma$  and  $\mu$  Lipschitz in  $x, y$  uniformly in  $t : \exists c > 0$  st  $\forall x, y \in \mathbb{R}, \forall t \in [a, b]$ 
  - $|\sigma(x, t) - \sigma(y, t)| \leq c|x - y|$
  - $|\mu(x, t) - \mu(y, t)| \leq c|x - y|$
- $\mathbb{E}(\xi_a^2) < \infty$

Then, there exists a solution  $X = (X_t)_{t \in [a, b]} \in M^2$  of the SDE, and it unique (up to un-discernibility).

*Proof for Theorem.*



### Theorem 11.1.3: Stochastic exponential

$$\begin{aligned}dX_t &= X_t dB_t \\ X_0 &= 1\end{aligned}$$

has for unique solution  $X_t = e^{B_t - \frac{t}{2}}$  (see next chapter).  
More generally, let  $Z_t$  be a Itô's process:

$$\begin{aligned}dZ_t &= a_t dt + b_t dB_t \\ Z_0 &= 0\end{aligned}$$

Then the stochastic differential equation

$$\begin{aligned}dX_t &= X_t dZ_t = a_t X_t dt + b_t X_t dB_t \\ X_0 &= 1\end{aligned}$$

has a unique solution called **the stochastic exponential of  $Z$** :

$$\mathcal{E}(Z)(t) = \exp\left(Z_t - \frac{1}{2} \int_0^t b_s^2 ds\right) \quad \forall t \in [0, T]$$

*Proof for Theorem.*

[5] p184

### Theorem 11.1.4: Numerical approximation of the solution of a SDE - Euler-Maruyama

Let  $X$  be a Itô's process solution of:

$$\begin{aligned}dX_t &= \mu(X_t, t)dt + \sigma(X_t, t)dB_t \\ X_a &= \xi_a\end{aligned}$$

where  $\mu(X_t, t)$  is the **drift** and  $\sigma^2(X_t, t)$  the **diffusion coefficient**.

Then one can approximate numerically the solution of the SDE by:

$$\begin{aligned}a &= t_0 < t_1 < \dots < t_k < t_{k+1} < \dots < t_n = b \\ x^{(k+1)} &= x^{(k)} + \mu(x^{(k)}, t_k)\Delta t + \sigma(x^{(k)}, t_k) \sqrt{\Delta t} \mathcal{N}(0, 1) \\ t_{k+1} &= t_k + \Delta t\end{aligned}$$

*Proof for Theorem.*

We have  $\Delta X_t \sim \mu(X_t, t)\Delta t + \sigma(X_t, t)\Delta B_t$ , where  $\Delta B_t \sim \mathcal{N}(0, \Delta t)$ . So  $\sigma(X_t, t)\Delta B_t \sim \mathcal{N}(0, \sigma^2(X_t, t)\Delta t)$ , and  $\Delta X_t \sim \mathcal{N}(\mu(X_t, t)\Delta t, \sigma^2(X_t, t)\Delta t)$

### SDE solutions are Markov Processes

Intuitively, when considering the solution  $X_t$  of 11.5 between  $t$  and  $t + \Delta t$ , leads to:

$$X_{t+\Delta t} = X_t + \sigma(X_t, t)\Delta B_t + \mu(X_t, t)\Delta t$$

We know that  $\Delta B_t$  is independent of  $\mathcal{F}_t$ , so  $X_{t+\Delta t}$  depends on the past only by  $X_t$ . This suggest  $(X_t)_{t \geq 0}$  is a Markov

Process.

We actually have the following theorem:

### Theorem 11.1.5: SDE solutions are Markov Processes

Consider the following SDE:

$$\begin{aligned} dX_t &= \mu(X_t, t)dt + \sigma(X_t, t)dB_t \\ X_0 &= \xi_0 \end{aligned}$$

where  $\sigma$  and  $\mu$  verify the hypothesis in 11.1. The solution  $X_t$  is given by:

$$X_t = \xi_0 + \int_0^t \mu(X_u, u)du + \int_0^t \sigma(X_u, u)dB_u$$

Then **X is Markov process**, with transition kernels given by ( $\forall x \in \mathbb{R}, \forall A \in \mathcal{B}_{\mathbb{R}}, \forall t \geq s$ ):

$$\begin{aligned} P_{s,t}(x, A) &= \mathbb{P}(X_t^{x,s} \in A) \\ X_t^{x,s} &= x + \int_s^t \mu(X_u^{x,s}, u)du + \int_s^t \sigma(X_u^{x,s}, u)dB_u \end{aligned}$$

ie  $X^{x,s} = (X_t^{x,s})_{t \geq s}$  is the solution of the SDE starting from  $x$  at time  $t$ .

If the SDE is time-invariant, ie

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t \quad (11.7)$$

$$X_0 = \xi_0 \quad (11.8)$$

then the Markov process solution  $X$  is homogeneous. (ie  $P_{s,t}$  depends only on  $t - s$ ). ie (for  $f$  Borelian bounded):

$$P_t f(x) = \int_{\mathbb{R}} f(z)P_t(x, dz) = \int_{\mathbb{R}} f(z)\mathbb{P}(X_t^{x,0} \in dz)$$

**Proof for Theorem.**

[5] p173

We can compute the infinitesimal generator when  $f$  is smooth enough:

### Theorem 11.1.6: Infinitesimal generator of $X$

In the case of time-invariant SDE 11.7 (when  $X$  is homogeneous), then : for any  $f \in C^2(\mathbb{R})$  bounded, with derivatives bounded, for any  $t > 0$ , for any  $x \in \mathbb{R}$ :

$$P_t f(x) = f(x) + \int_0^t P_s(Af)(x)ds \quad (11.9)$$

$$Af(x) = \frac{1}{2}\sigma^2(x)f''(x) + \mu(x)f'(x) \quad (11.10)$$

**Proof for Theorem.**

When the probability transitions have densities, we end up with the Kolmogorov equations:

### Theorem 11.1.7: Kolmogorov equations for a general SDE

Consider SDE:

$$\begin{aligned}dX_t &= \mu(X_t, t)dt + \sigma(X_t, t)dB_t \\ X_0 &= \xi_0\end{aligned}$$

We assume the transition probabilities have densities:

$$P_t(x, dy) = p_t(x, y)dy \quad (x, y) \in \mathbb{R}$$

Then, remembering that  $x$  is the "start" and  $y$  the "arrival":

- $\frac{\partial}{\partial t} p_t(x, y) = \left( \frac{1}{2} \sigma(x)^2 \frac{\partial^2}{\partial x^2} + \mu(x) \frac{\partial}{\partial x} \right) p_t(x, y)$
- $\frac{\partial}{\partial t} p_t(x, y) = \frac{1}{2} \frac{\partial^2}{\partial y^2} (\sigma(y)^2 p_t(x, y)) - \frac{\partial}{\partial y} (\mu(y) p_t(x, y))$

The second equation ("futur Kolmogorov") is also known as the Fokker-Plank equation.

***Proof for Theorem.***

## Quantifying randomness with Entropy

In its most general form, a data sequence can be considered as a realization of a stochastic process. A stochastic process is usually described as either a sequence of random variables (ie a set of  $(X_t)_{t \geq 0} : \omega \rightarrow X_t(\omega)$ ), or a single random variable over the space of functions (ie  $X : \omega \rightarrow X(\omega) = X_\omega = \{t \rightarrow X_\omega(t)\}$ ).

As we wish to measure -somehow- the degree of "randomness" of a data sequence, we will find more convenient to use the former view (ie consider a data sequence as a countable sequence of random variables), as it allows to use the framework of information theory.

First, we will recall the basic definitions of information theory : entropy, relative entropy (ie KL divergence), conditional entropy and mutual information. Then we will write some results regarding the application of Information Theory (IT) to data sequence. Last, we will describe two of the most used empirical measurements : Approximate Entropy (ApEn) and Sample Entropy (SampEn).

The basic of IT are introduced in [8], [17], and [15]. One of the reference books on the subject is [4], and goes much further than the scope of this report. Of course, the interested reader will also refer to the seminal paper by Shannon : [1].

**Entropy** : given a random variable  $X$ , either discrete or continuous, taking values in a measurable space  $\mathcal{X}$ , and its probability distribution  $p$ , the amount of information given by a given realization  $x$  is given by  $\log \frac{1}{p(x)} = -\log p(x)$  (the lower the probability, the higher the amount of information).

The average amount of information (over all possible values of  $x$ ) required to describe the random variable  $X$  is the entropy of  $X$  :

$$\mathcal{H}(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x) \quad (12.1)$$

(or  $-\int_{\mathcal{X}} p(x) \log p(x) dx$ ).

**Relative entropy, KL divergence** : when approximating the true data distribution  $p$  by a distribution  $q$ , we require in average a quantity of information  $-\sum_{x \in \mathcal{X}} p(x) \log q(x)$  to describe  $X$ . The difference between the optimal amount of information (ie the entropy  $\mathcal{H}(X)$ ) and this quantity is the well-known relative entropy, of KL-divergence between

$p$  and  $q$  :

$$\mathbb{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \quad (12.2)$$

$$= \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx \quad (12.3)$$

The properties of KL-divergence (positiveness, non symmetry) are well described in the sources above.

**Conditional Entropy** : we now consider two random variables  $X$  and  $Y$ , and wish to measure the degree of relationship between them. We define the conditional entropy of, for example,  $Y$  given  $X$ , as the amount of information we get observing the values of  $Y$  given  $X$ , averaged over the joint probability. Formally :

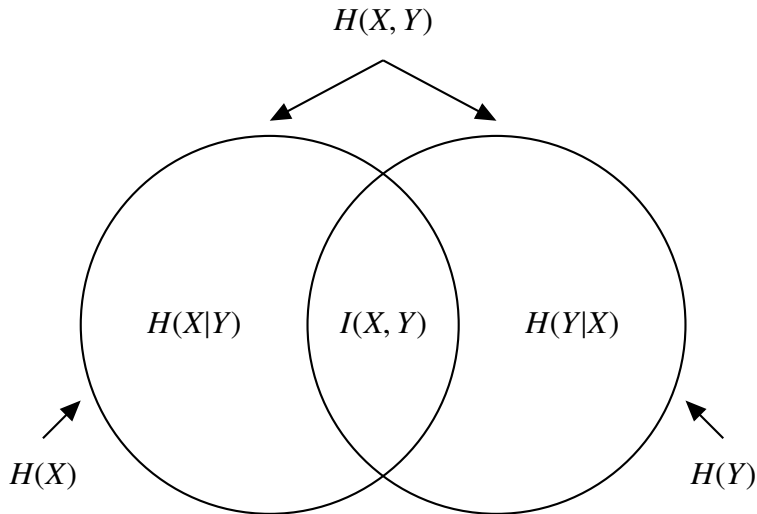
$$\begin{aligned} \mathcal{H}(Y|X) &= - \sum_{x,y \in \mathcal{X}, \mathcal{Y}} p(x,y) \log p(y|x) \\ &= - \int_{\mathcal{X}, \mathcal{Y}} p(x,y) \log p(y|x) dx dy \end{aligned}$$

By basic calculation, we get  $\mathcal{H}(X, Y) = \mathcal{H}(Y|X) + \mathcal{H}(X) = \mathcal{H}(X|Y) + \mathcal{H}(Y)$

**Mutual Information** - last, still considering two random variables  $X, Y$ , the mutual information is the additional amount of information we need to describe  $X, Y$  when we assume independence (ie use  $p(x)p(y)$ ) rather than use the true joint probability  $p(x, y)$ . This amount is  $-\log p(x)p(y) - (-\log p(x, y)) = \log \frac{p(x, y)}{p(x)p(y)}$ , that we average over the true distribution  $p(x, y)$ :

$$\mathcal{I}(X, Y) = \sum_{x,y \in \mathcal{X}, \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (12.4)$$

The relationships between entropy, relative entropy, conditional entropy and mutual information, are well described using a Venn diagram:



For example :  $\mathcal{I}(X, Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X, Y)$ , etc.

If we now consider a sequence of  $n$  random variables, a way to describe the randomness of the sequence is to measure how the entropy of the joint distribution grow with  $n$ . (see [4] p63). Typically, if the random variables are independent, we can expect the entropy to grow at each step  $n$  by the full amount of  $\mathcal{H}(X_n)$ . On the other hand, if

correlations exist, then we can expect the overall entropy to grow by a lesser amount.

Formally, the **entropy rate** of a stochastic process  $(X_n)_{n \in \mathbb{N}^*}$  is defined by:

$$\mathcal{H}(X) = \lim_{n \rightarrow \infty} \frac{1}{n} \mathcal{H}(X_1, X_2, \dots, X_n) \quad (12.5)$$

when the limit exists. This is the entropy per symbol.

[4] also defines a **conditional entropy rate**

$$\mathcal{H}'(X) = \lim_{n \rightarrow \infty} \frac{1}{n} \mathcal{H}(X_n | X_{n-1}, \dots, X_2, X_1) \quad (12.6)$$

when the limit exists. This is the entropy of the latest random variable, given the past realizations.

An important result is that, when  $(X_n)$  is stationary, both limits exist and are equal.

The question naturally arises that, given a data sequence, how do we compute an approximation of 12.5 or 12.6. We now introduce the ApEn and SampEn (see [10] and [2]).

**Approximate Entropy** - Approximate Entropy is a measure of the log probability that patterns, that were identified in the data through the examination of sub sequences of a given length, still remain when considering longer subsequences.

Formally, let  $x = (x_1, x_2, \dots, x_N)$  be a data sequence of length  $N$ ,  $m$  an integer ( $0 < m \leq N$ ), and  $r > 0$  a measure of acceptable noise. We define as *blocks* of length  $m$  the subsequences  $b_i^m = \{x_i, x_{i+1}, \dots, x_{i+m-1}\}$ , starting at  $i$  with ( $1 \leq i \leq N - m + 1$ ). For two blocks  $b_i^m$  and  $b_j^m$ , we define a component-wise distance  $d_{ij}^m = \max_{k=0,1,\dots,m-1} |b_{i+k}^m - b_{j+k}^m|$ .

We consider "close enough" (ie similar) those blocks whose distance is lower than the acceptable noise (tolerance)  $r$ , and calculate the frequency of those similar blocks ( $b_i^m, b_j^m$ ) w.r.t. all blocks of length  $m$  for a given  $b_i^m$ :

$$C_i^m(r) = \frac{\text{number of } j \leq N - m + 1 \text{ s.t. } d_{ij} \leq r}{\text{number of blocks of length } m : N - m + 1} \quad (12.7)$$

We then average the  $\log C_i^m$  over all possible subsequences  $b_i$ :

$$\Phi^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \log C_i^m(r)$$

We can see  $\Phi^m(r)$  as the average of the log probability of two subsequences of length  $m$  to be similar (up to the tolerance  $r$ ). Finally, we compute:

$$\text{ApEn}(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r) \quad (12.8)$$

When  $m \ll N$ , then  $-\text{ApEn}(m, r, N) \sim \frac{1}{N+1} \sum_{i=1}^{N+1} \log \frac{C_i^{m+1}}{C_i^m}$ , which is the average over  $i$  of the log of the conditional probability of two sequences  $b_i^{m+1}, b_j^{m+1}$  of length  $m+1$  be similar given that they are already similar for lengths  $1, 2, \dots, m$ .

Last,  $\text{ApEn}(m, r, N)$  is a statistical estimator of:

$$\text{ApEn}(m, r) = \lim_{N \rightarrow \infty} \text{ApEn}(m, r, N) \quad (12.9)$$

Intuitively, the greater the regularity in a data sequence, the greater the likelihood that patterns existing for subsequences of length  $m$  still remain for subsequences of greater length, ie the smaller ApEn, and conversely.

Key properties of ApEn include:

- ApEn is independent of any model of the data sequence.
- due to its construction, ApEn is non-negative, is finite for stochastic processes and deterministic processes with noise.
- following [2], it is imperative to eliminate any trend in the data sequence before computing ApEn and drawing conclusions.
- typical recommended values for  $m$  are 2 and 3. Typical recommended values for  $r$  are in the range of 0.1 to 0.25 the standard deviation of the data sequence, in order to allow a sufficient number of subsequences close within a distance  $r$ , and reasonable estimates of the conditional probabilities.
- if the noise is significant (ie signal-to-noise ratio lower than three), then precautions must be taken with the interpretation of ApEn.
- Regular measurements of the data sequence over time are required.
- Normalization of data sequences is required before computations of ApEn to compare data sequences between each other.

**Sample Entropy** - we can see in 12.7 that a given subsequence  $b_i^m$  is counted in both the numerator and the denominator. If this ensures numerical stability (ie no attempt to calculate  $\log 0$  for example), this also introduces a bias in the calculation of the probability estimates. SampEn explicitly discounts the subsequence  $b_i^m$  from the calculations to remove the bias.

Formally:

$$A_i^m(r) = \frac{1}{N - m - 1} \{\text{number of blocks } b_j^{m+1} \text{ of length } m + 1 \text{ s.t. } i \neq j \text{ and } d_{ij}^{m+1} \leq r\} \quad (12.10)$$

$$B_i^m(r) = \frac{1}{N - m} \{\text{number of blocks } b_j^m \text{ of length } m \text{ s.t. } i \neq j \text{ and } d_{ij}^m \leq r\} \quad (12.11)$$

$$B^m(r) = \frac{1}{N - m} \sum_{i=1}^{N-m} B_i^m(r) \quad (12.12)$$

$$A^m(r) = \frac{1}{N - m - 1} \sum_{i=1}^{N-m-1} A_i^m(r) \quad (12.13)$$

And

$$\text{SampEn}(m, r, N) = -\log \frac{A^m(r)}{B^m(r)} \quad (12.14)$$

$$\text{SampEn}(m, r) = -\lim_{n \rightarrow \infty} \log \frac{A^m(r)}{B^m(r)} \quad (12.15)$$

## Perspective 1 : Linear SDEs and Markovian kernel GPs

Where we discuss stuff.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.



## Perspective 2 : more IT on stochastic processes

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

# **Part IV**

## **Experiments**

## XPs

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

## **Part V**

# **Conclusion and Discussion**

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

# **Part VI**

## **Appendices**

# **Appendices**

## Vanilla Variational Auto Encoder

We consider a sequence of i.i.d points  $(x_i)_{i=1,\dots,N} \in \mathbb{R}^D$ , and the associated latent variables  $(z_i)_{i=1,\dots,N} \in \mathbb{R}^L$ .

In the vanilla VAE setting, the observation model (decoder) is  $p_{\theta_x}(x|z)$ , the approximate posterior (encoder) is  $q_{\phi}(z|x)$ , the latent prior is  $p_{\theta_z}(z)$ .

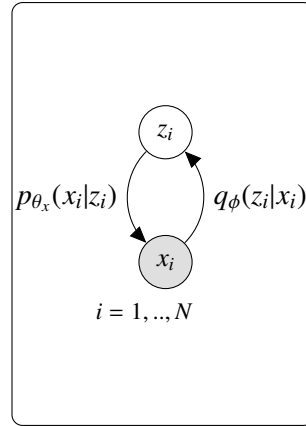


Figure A.1: Vanilla VAE

The log likelihood of the data is:

$$\log p_{\theta}(x) = \log \frac{p(x, z)}{p(z|x)}$$

Multiplying both sides by  $q_{\phi}(z|x)$  and integrating over  $dz$  leads to:

$$\begin{aligned} \log p_{\theta}(x) &= \int q_{\phi}(z|x) \log \frac{p_{\theta}(x, z)}{p(z|x)} dz \\ &= \int q_{\phi}(z|x) \log \frac{p_{\theta}(x, z)}{q_{\phi}(z|x)} \frac{q_{\phi}(z|x)}{p(z|x)} dz \\ &= \mathbb{E}_{q_{\phi}(z|x)} \log \frac{p_{\theta}(x, z)}{q_{\phi}(z|x)} + \mathbb{KL}(q_{\phi}(z|x) || p(z|x)) \\ &\geq \mathbb{E}_{q_{\phi}(z|x)} \log \frac{p_{\theta}(x, z)}{q_{\phi}(z|x)} = \mathcal{L}(\theta, \phi, X) \end{aligned}$$



In this setting, the D-separation is obvious and the joint distribution factorizes over  $n$ :

$$p_{\theta}(x, z) = \prod_{i=1}^n p_{\theta_x}(x_i|z_i) p_{\theta_z}(z_i)$$

$$q_{\phi}(z|x) = \prod_{i=1}^n q_{\phi}(z_i|x_i)$$

The VLB (or ELBO)  $\mathcal{L}(\theta, \phi, X)$  simplifies into:

$$\begin{aligned} \mathcal{L}(\theta, \phi, X) &= \mathbb{E}_{q_{\phi}(z|x)} \log \frac{\prod_{i=1}^n p_{\theta_x}(x_i|z_i) p_{\theta_z}(z_i)}{\prod_{i=1}^n q_{\phi}(z_i|x_i)} \\ &= \sum_{i=1}^n \mathbb{E}_{q_{\phi}(z_i|x_i)} p_{\theta_x}(x_i|z_i) - \sum_{i=1}^n \mathbb{KL}(q_{\phi}(z_i|x_i) || p_{\theta_z}(z_i)) \end{aligned}$$

The first term is the reconstruction loss, and is estimated via Monte Carlo sampling over  $z_i \sim q_{\phi}(z_i|x_i)$ . The second term is a KL-divergence, which can be computed analytically when  $q_{\phi}$  and  $p_{\theta_z}$  are chosen to be Gaussians.

## Gaussian Process

We summarize here most of the results of the Gaussian Process, and refers the reader to [6] for further details.

We first recall the Gaussian marginal and conditional result:

Let  $x$  and  $y$  be jointly Gaussian vectors, ie:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \tilde{A} & \tilde{C} \\ \tilde{C}^T & \tilde{B} \end{bmatrix}^{-1}\right) \quad (\text{B.1})$$

where  $A, B, C$  is the block decomposition of the covariance matrix, and  $\tilde{A}, \tilde{B}, \tilde{C}$  the block decomposition of the precision matrix.

Then the marginal distribution of  $x$  and the conditional distribution of  $x$  given  $y$  are :

$$x \sim \mathcal{N}(\mu_x, A) \quad (\text{B.2})$$

$$x|y \sim \mathcal{N}(\mu_x + CB^{-1}(y - \mu_y), A - CB^{-1}C^T) \quad (\text{B.3})$$

$$= \mathcal{N}(\mu_x - \tilde{A}^{-1}\tilde{C}(y - \mu_y), \tilde{A}^{-1}) \quad (\text{B.4})$$

We now consider a Gaussian Process with mean function  $m(\cdot)$  and kernel  $k(\cdot, \cdot)$

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \quad (\text{B.5})$$

At the training points  $X = \{x_1, \dots, x_n\}$ , the observations are  $Y = \{y_1, \dots, y_n\}$  with some noise  $y = f(x) + \epsilon$  with  $\epsilon \stackrel{i.i.d}{\sim} \mathcal{N}(0, \sigma_n^2)$ .

The covariance between observations writes:

$$\text{cov}(y_p, y_q) = k(x_p, x_q) + \delta_{pd}\sigma_n^2 \quad (\text{B.6})$$

$$\text{cov}(y) = K(X, X) + \sigma_n^2 I \quad (\text{B.7})$$

At some test points  $X_*$ , we aim to predict  $f_* = f(X_*)$ . Then:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right) \quad (\text{B.8})$$

From which we get:

$$f_*|X_*, X, Y \sim \mathcal{N}(\bar{f}_*, \text{cov}(f_*)) \quad (\text{B.9})$$

$$\bar{f}_* = K(X_*, X) \left( K(X, X) + \sigma_n^2 I \right)^{-1} Y \quad (\text{B.10})$$

$$\text{cov}(f_*) = K(X_*, X_*) - K(X_*, X) \left( K(X, X) + \sigma_n^2 I \right)^{-1} K(X, X_*) \quad (\text{B.11})$$

## KL divergence between two exponential-family distributions

We recall the family of distributions parameterized by  $\eta \in \mathbb{R}^K$ , over a fixed support  $\mathcal{X}^D \in \mathbb{R}^D$  : the **exponential family** of distributions  $p(x|\eta)$  is given by:

$$p(x|\eta) = \frac{1}{Z(\eta)} h(x) \exp(\eta^T \mathcal{T}(x)) \quad (\text{C.1})$$

$$= h(x) \exp(\eta^T \mathcal{T}(x) - A(\eta)) \quad (\text{C.2})$$

with:

- $h(x)$  is the base measure, ie a scaling constant (often 1)
- $\mathcal{T}(x)$  are the sufficient statistics
- $\eta$  are the natural parameters, or canonical parameters
- $Z(\eta)$  is the partition function,  $A(\eta)$  is the log partition function.

The Bernoulli, categorical (ie multinomial for one observation), Gaussian distributions are part of the exponential family.

The **KL-divergence between two exponential family distributions of the same family** is:

$$\mathbb{KL}(p(x|\eta_1)||p(x|\eta_2)) = \mathbb{E}_{\eta_1} [(\eta_1 - \eta_2)\mathcal{T}(x) - A(\eta_1) + A(\eta_2)] \quad (\text{C.3})$$

$$= (\eta_1 - \eta_2)^T \mathbb{E}_{\eta_1} \mathcal{T}(x) - A(\eta_1) + A(\eta_2) \quad (\text{C.4})$$

The most important example is the **KL-divergence between two multivariate Gaussian distributions of dimension  $D$** :

**KL between two multivariate Gaussians of dimension  $D$**

$$\mathbb{KL}(\mathcal{N}(x|\mu_1, \Sigma_1)||\mathcal{N}(x|\mu_2, \Sigma_2)) = \frac{1}{2} \left[ \text{tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) - D + \log \frac{|\Sigma_2|}{|\Sigma_1|} \right] \quad (\text{C.5})$$



## Bibliography

- [1] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948, issn: 0005-8580. doi: [10.1002/j.1538-7305.1948.tb01338.x](https://doi.org/10.1002/j.1538-7305.1948.tb01338.x) Accessed: Jul. 28, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/6773024>
- [2] S. M. Pincus, “Approximate entropy as a measure of system complexity,” en, *Proceedings of the National Academy of Sciences*, vol. 88, no. 6, pp. 2297–2301, Mar. 1991, issn: 0027-8424, 1091-6490. doi: [10.1073/pnas.88.6.2297](https://doi.org/10.1073/pnas.88.6.2297) Accessed: Jan. 20, 2025. [Online]. Available: <https://pnas.org/doi/full/10.1073/pnas.88.6.2297>
- [3] C. Bishop, *Pattern recognition and machine learning*, Accessed on Month Day, Year, 2006. [Online]. Available: <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>
- [4] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition*, English. Hoboken, N.J: Wiley-Interscience, 2006, isbn: 9780471241959.
- [5] *Mouvement brownien et calcul d’Itô-Léonard Gallardo-Editions Hermann*, Sep. 2008. Accessed: Apr. 16, 2025. [Online]. Available: <https://www.editions-hermann.fr/livre/mouvement-brownien-et-calcul-d-ito-leonard-gallardo>
- [6] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning* (Adaptive computation and machine learning), eng, 3. print. Cambridge, Mass.: MIT Press, 2008, isbn: 9780262182539.
- [7] S. Roberts, M. Osborne, M. Ebdon, S. Reece, N. Gibson, and S. Aigrain, “Gaussian processes for time-series modelling,” en, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, p. 20110550, Feb. 2013, issn: 1364-503X, 1471-2962. doi: [10.1098/rsta.2011.0550](https://doi.org/10.1098/rsta.2011.0550) Accessed: Jul. 23, 2025. [Online]. Available: <https://royalsocietypublishing.org/doi/10.1098/rsta.2011.0550>
- [8] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information science and statistics), eng, Soft-cover reprint of the original 1st edition 2006 (corrected at 8th printing 2009). New York, NY: Springer New York, 2016, isbn: 9781493938438.
- [9] F. P. Casale, A. V. Dalca, L. Saglietti, J. Listgarten, and N. Fusi, *Gaussian Process Prior Variational Autoencoders*, en, Oct. 2018. Accessed: Jun. 7, 2025. [Online]. Available: <https://arxiv.org/abs/1810.11738v2>
- [10] A. Delgado-Bonal and A. Marshak, “Approximate Entropy and Sample Entropy: A Comprehensive Tutorial,” en, *Entropy*, vol. 21, no. 6, p. 541, Jun. 2019, issn: 1099-4300. doi: [10.3390/e21060541](https://doi.org/10.3390/e21060541) Accessed: Jan. 20, 2025. [Online]. Available: <https://www.mdpi.com/1099-4300/21/6/541>
- [11] D. P. Kingma and M. Welling, “An Introduction to Variational Autoencoders,” 2019. doi: [10.48550/ARXIV.1906.02691](https://doi.org/10.48550/ARXIV.1906.02691) Accessed: Jul. 21, 2025. [Online]. Available: <https://arxiv.org/abs/1906.02691>

- [12] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations* (Institute of Mathematical Statistics Textbooks). Cambridge: Cambridge University Press, 2019, ISBN: 9781316510087. DOI: [10.1017/9781108186735](https://doi.org/10.1017/9781108186735) Accessed: Jul. 23, 2025. [Online]. Available: <https://www.cambridge.org/core/books/applied-stochastic-differential-equations/6BB1B8B0819F8C12616E4A0C78C29EAA>
- [13] V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt, *GP-VAE: Deep Probabilistic Time Series Imputation*, arXiv:1907.04155, Feb. 2020. DOI: [10.48550/arXiv.1907.04155](https://doi.org/10.48550/arXiv.1907.04155) Accessed: Jun. 7, 2025. [Online]. Available: <http://arxiv.org/abs/1907.04155>
- [14] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, *Dynamical Variational Autoencoders: A Comprehensive Review*, arXiv:2008.12595, Jul. 2022. DOI: [10.48550/arXiv.2008.12595](https://doi.org/10.48550/arXiv.2008.12595) Accessed: Jan. 19, 2025. [Online]. Available: <http://arxiv.org/abs/2008.12595>
- [15] K. Murphy, *Probabilistic machine learning advanced topics*, 2023. [Online]. Available: <https://mitpress.mit.edu/9780262048439/probabilistic-machine-learning/>
- [16] H. Zhu, C. B. Rodas, and Y. Li, *Markovian Gaussian Process Variational Autoencoders*, arXiv:2207.05543, Aug. 2023. DOI: [10.48550/arXiv.2207.05543](https://doi.org/10.48550/arXiv.2207.05543) Accessed: Jul. 23, 2025. [Online]. Available: <http://arxiv.org/abs/2207.05543>
- [17] F. Koller, *Probabilistic Graphical Models*, en-US. Accessed: Jul. 21, 2025. [Online]. Available: <https://mitpress.mit.edu/9780262013192/probabilistic-graphical-models/>
- [18] *Page Web de Jean-François Le Gall*. Accessed: Apr. 16, 2025. [Online]. Available: <https://www.imo.universite-paris-saclay.fr/~jean-francois.le-gall/>