Variational autoencoder with weighted samples for high-dimensional non-parametric adaptive importance sampling

Julien Demange-Chryst, PhD student

julien.demange-chryst@onera.fr

Co-authors: Jérôme Morio[1], François Bachoc[2], Timothé Krauth[1,3]

[1]DTIS, ONERA, Université de Toulouse, 31000 Toulouse France, [2]Institut de Mathématiques de Toulouse, University Toulouse III, [3]Zurich University of Applied Sciences, Centre for Aviation

Outline

Outline

Context
Uncertainty quantification

Numerical code

$\psi : \mathbb{X} \subseteq \mathbb{R}^d \longrightarrow \mathbb{R}$

Characteristics of the numerical code ψ:

- black-box model
- deterministic
- expensive to evaluate
  $\hookrightarrow$ cost of an algorithm: number of calls to ψ

Context

Uncertainty quantification

Inputs
$\mathbf{X} = (X_1, \ldots, X_d)^\top \sim f_{\mathbf{X}}$

→

Numerical code
$\psi : \mathbb{X} \subseteq \mathbb{R}^d \longrightarrow \mathbb{R}$

Characteristics of the random vector $\mathbf{X}$:

- $f_{\mathbf{X}}$ d-dimensional continuous distribution
- $f_{\mathbf{X}}$ fully known
- potentially with dependent components

Characteristics of the numerical code ψ:

- black-box model
- deterministic
- expensive to evaluate
  $\hookrightarrow$ cost of an algorithm: number of calls to ψ

Context

Uncertainty quantification

Uncertainty propagation $\Longrightarrow$

Inputs $\mathbf{X} = (X_1, \dots, X_d)^\top \sim f_{\mathbf{X}}$ → **Numerical code** $\psi : \mathbb{X} \subseteq \mathbb{R}^d \longrightarrow \mathbb{R}$ → **Output** $Y = \psi(\mathbf{X}) \in \mathbb{R}$

Characteristics of the random vector $\mathbf{X}$:

- $f_{\mathbf{X}}$ d-dimensional continuous distribution
- $f_{\mathbf{X}}$ fully known
- potentially with dependent components

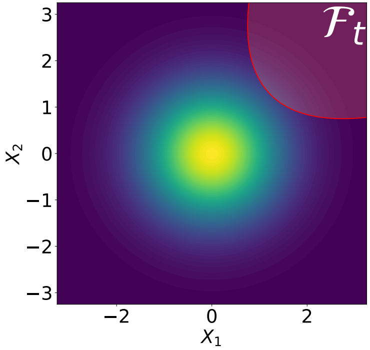
Characteristics of the numerical code ψ:

- black-box model
- deterministic
- expensive to evaluate
  $\hookrightarrow$ cost of an algorithm: number of calls to ψ

Context

Reliability analysis

Inputs $\mathbf{X} = (X_1, \dots, X_d)^\top \sim f_{\mathbf{X}}$ → **Numerical code** $\psi : \mathbb{X} \subseteq \mathbb{R}^d \longrightarrow \mathbb{R}$ → **Output** $Y = \psi(\mathbf{X}) \in \mathbb{R}$ → **Quantity of interest** $\mathbb{1}(\psi(\mathbf{X}) > t)$



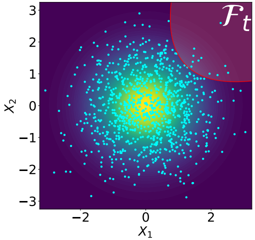
- $t \in \mathbb{R}$ is a **critical threshold**
- $\{\psi(\mathbf{X}) > t\}$ is the **failure event**
- the **failure domain** is $\mathcal{F}_t = \{\mathbf{x} \in \mathbb{X} \,/\, \psi(\mathbf{x}) > t\}$
- the **limit state** is $\{\mathbf{x} \in \mathbb{X} \,/\, \psi(\mathbf{x}) = t\}$

Failure probability:

$$p_t = \mathbb{P}(\psi(\mathbf{X}) > t) = \int_{\mathcal{F}_t} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} = \mathbb{E}_{f_{\mathbf{X}}}\left[\mathbb{1}(\psi(\mathbf{X}) > t)\right]$$

Rare event estimation

Crude Monte Carlo method and alternatives



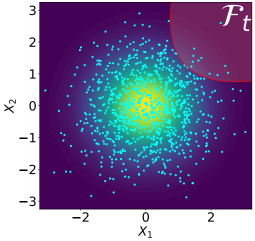
Classical crude Monte Carlo method:

$$\widehat{p}_{t,N}^{\mathrm{MC}} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(\psi\left(\mathbf{X}^{(n)}\right) > t\right) \text{ with } \left(\mathbf{X}^{(n)}\right)_{n\in[\![1,N]\!]} \sim f_{\mathbf{X}}$$

✘ if $p_t \approx 10^{-a}$, we need $N \approx 10^{a+2}$ to have an error of 10%

Rare event estimation

Crude Monte Carlo method and alternatives



Classical crude Monte Carlo method:

$$\widehat{p}_{t,N}^{\mathrm{MC}} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(\psi\left(\mathbf{X}^{(n)}\right) > t\right) \text{ with } \left(\mathbf{X}^{(n)}\right)_{n\in[\![1,N]\!]} \sim f_{\mathbf{X}}$$

✘ if $p_t \approx 10^{-a}$, we need $N \approx 10^{a+2}$ to have an error of 10%

Other existing methods:

- deterministic methods such as FORM/SORM [HL74, Bre84]
- subset sampling [CDMFG12]
- importance sampling [Buc04]

Principle of importance sampling

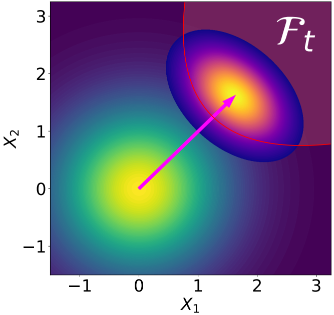
Consider an auxiliary sampling distribution g to draw more samples in $\mathcal{F}_t$ than $f_{\mathbf{X}}$

Rare event estimation

Importance sampling

Principle of importance sampling

Consider an auxiliary sampling distribution g to draw more samples in $\mathcal{F}_t$ than $f_{\mathbf{X}}$

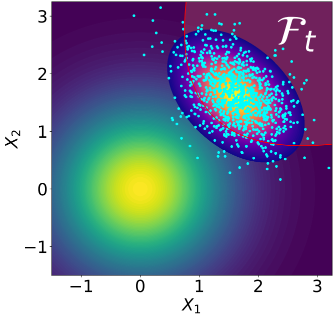


Rewriting p_t according to g:

$$p_t = \mathbb{E}_{f_{\mathbf{X}}}\left[\mathbb{1}\left(\psi\left(\mathbf{X}\right) > t\right)\right] = \mathbb{E}_g\left[\mathbb{1}\left(\psi\left(\mathbf{X}\right) > t\right)\frac{f_{\mathbf{X}}\left(\mathbf{X}\right)}{g\left(\mathbf{X}\right)}\right]$$

Principle of importance sampling

Consider an auxiliary sampling distribution g to draw more samples in $\mathcal{F}_t$ than $f_{\mathbf{X}}$



Rewriting p_t according to g:

$$p_t = \mathbb{E}_{f_{\mathbf{X}}}\left[\mathbb{1}\left(\psi\left(\mathbf{X}\right) > t\right)\right] = \mathbb{E}_g\left[\mathbb{1}\left(\psi\left(\mathbf{X}\right) > t\right)\frac{f_{\mathbf{X}}\left(\mathbf{X}\right)}{g\left(\mathbf{X}\right)}\right]$$

Importance sampling estimator of p_t:

$$\widehat{p}^{\mathrm{IS}}_{t,N} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}\left(\psi\left(\mathbf{X}^{(n)}\right) > t\right)\frac{f_{\mathbf{X}}\left(\mathbf{X}^{(n)}\right)}{g\left(\mathbf{X}^{(n)}\right)}$$

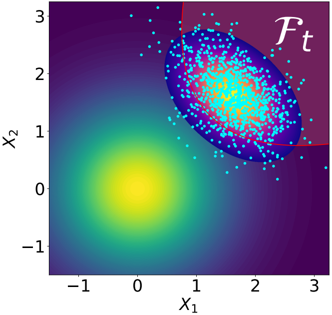
with $\left(\mathbf{X}^{(n)}\right)_{n\in[\![1,N]\!]} \sim g$

Rare event estimation

Importance sampling

Principle of importance sampling

Consider an auxiliary sampling distribution g to draw more samples in $\mathcal{F}_t$ than $f_{\mathbf{X}}$



Rewriting p_t according to g:

$$p_t = \mathbb{E}_{f_{\mathbf{X}}}\left[\mathbb{1}\left(\psi\left(\mathbf{X}\right) > t\right)\right] = \mathbb{E}_g\left[\mathbb{1}\left(\psi\left(\mathbf{X}\right) > t\right)\frac{f_{\mathbf{X}}\left(\mathbf{X}\right)}{g\left(\mathbf{X}\right)}\right]$$

Optimal IS auxiliary distribution [Buc04]:

$$g_{\text{opt}}\left(\mathbf{x}\right) = \frac{\mathbb{1}\left(\psi\left(\mathbf{x}\right) > t\right) f_{\mathbf{X}}\left(\mathbf{x}\right)}{p_t} = f_{\mathbf{X}|\mathbf{X}\in\mathcal{F}_t}\left(\mathbf{x}\right)$$

$\Longrightarrow$ in practice, g_{opt} is approximated

Importance Sampling

Approximation of the optimal auxiliary distribution

Question: How do we approximate g_{opt}?

- within a parametric family (ex: Gaussian [RK04], Gaussian mixture [GPS19])
- by a non-parametric model (ex: kernel smoothing) [Zha96, Mor11, FCIM23]

Importance Sampling

Approximation of the optimal auxiliary distribution

Question: How do we approximate g_{opt}?

- within a parametric family (ex: Gaussian [RK04], Gaussian mixture [GPS19])
- by a non-parametric model (ex: kernel smoothing) [Zha96, Mor11, FCIM23]

Robustness faced to the dimension

Flexibility

Importance Sampling

Approximation of the optimal auxiliary distribution

Question: How do we approximate g_{opt}?

- within a parametric family (ex: Gaussian [RK04], Gaussian mixture [GPS19])
- by a non-parametric model (ex: kernel smoothing) [Zha96, Mor11, FCIM23]

Robustness faced to the dimension

Flexibility

Main question

Is it possible to approximate g_{opt} by satisfying both characteristics?

Outline

Dimensionality reduction

Principle of dimensionality reduction

Reduce the number of features to describe and represent high dimensional data

Dimensionality reduction

Introduction

Principle of dimensionality reduction

Reduce the number of features to describe and represent high dimensional data

Methods to do so:

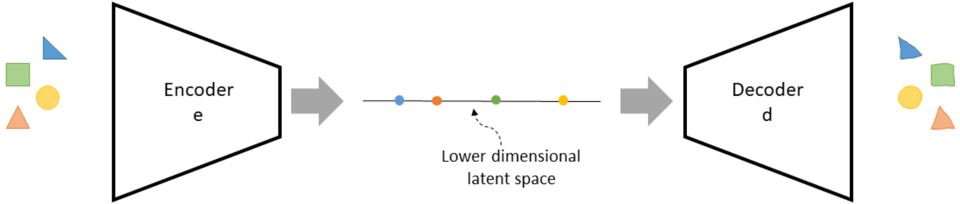
- selection: select a reduced number of existing features
- extraction: create a reduced number of new features based on the existing ones

Dimensionality reduction

Introduction

Principle of dimensionality reduction

Reduce the number of features to describe and represent high dimensional data



Examples:

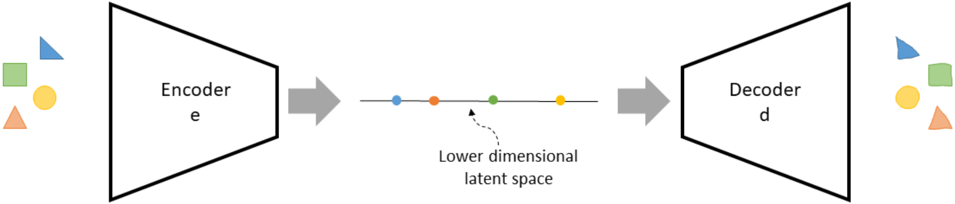
- PCA [WEG87]: encoder and decoder are linear transformations of the input data
- autoencoder [MRG+87]: encoder and decoder neural networks

Dimensionality reduction

Introduction

Principle of dimensionality reduction

Reduce the number of features to describe and represent high dimensional data



In that setting:

- ✔ encoding data into a lower dimensional latent space
- ✘ bad generation properties

Variational autoencoder

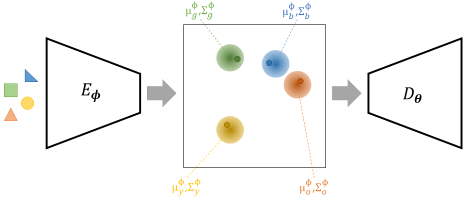
General presentation

A **variational autoencoder** (VAE) [KW14] can be seen as a regularised autoencoder

Variational autoencoder

General presentation

A **variational autoencoder** (VAE) [KW14] can be seen as a regularised autoencoder



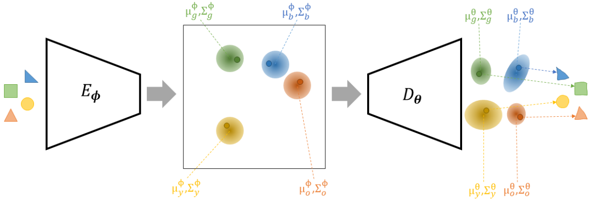
- an input data is encoded as a distribution

$$g_{\phi}(.|\mathbf{x}) = E_{\phi}(\mathbf{x}) = \mathcal{N}_{d_z}\left(\boldsymbol{\mu}_{\mathbf{x}}^{\phi}, \boldsymbol{\Sigma}_{\mathbf{x}}^{\phi}\right)$$

Variational autoencoder

General presentation

A **variational autoencoder** (VAE) [KW14] can be seen as a regularised autoencoder



- an input data is encoded as a distribution

$$g_\phi(.|\mathbf{x}) = E_\phi(\mathbf{x}) = \mathcal{N}_{d_z}\left(\boldsymbol{\mu}_\mathbf{x}^\phi, \boldsymbol{\Sigma}_\mathbf{x}^\phi\right)$$

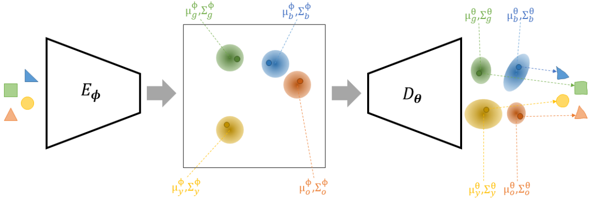
- a latent point is decoded as a distribution

$$g_\theta(.|\mathbf{z}) = D_\theta(\mathbf{z}) = \mathcal{N}_d\left(\boldsymbol{\mu}_\mathbf{z}^\theta, \boldsymbol{\Sigma}_\mathbf{z}^\theta\right)$$

Variational autoencoder

General presentation

A **variational autoencoder** (VAE) [KW14] can be seen as a regularised autoencoder



- an input data is encoded as a distribution

$$g_\phi(.|\mathbf{x}) = E_\phi(\mathbf{x}) = \mathcal{N}_{d_z}\left(\boldsymbol{\mu}_\mathbf{x}^\phi, \boldsymbol{\Sigma}_\mathbf{x}^\phi\right)$$

- a latent point is decoded as a distribution

$$g_\theta(.|\mathbf{z}) = D_\theta(\mathbf{z}) = \mathcal{N}_d\left(\boldsymbol{\mu}_\mathbf{z}^\theta, \boldsymbol{\Sigma}_\mathbf{z}^\theta\right)$$

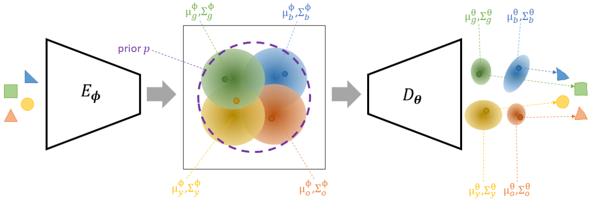
Loss function:

$$\underset{\boldsymbol{\phi},\boldsymbol{\theta}}{\arg\max}\ \underbrace{\mathbb{E}_{f_\mathbf{X}}\left[\mathbb{E}_{g_\phi(.|\mathbf{X})}\left(\log\left(g_\theta\left(\mathbf{X}|\mathbf{Z}\right)\right)\right)\right]}_{\text{log-likelihood}} - \dots$$

Variational autoencoder

General presentation

A **variational autoencoder** (VAE) [KW14] can be seen as a regularised autoencoder



Add a regularisation term to a prior p to:

- bring continuity and completeness to the latent space
- ✔ have good generation properties!

Loss function:

$$\underset{\phi,\theta}{\arg\max} \; \underbrace{\mathbb{E}_{f_{\mathbf{X}}}\left[\mathbb{E}_{g_{\phi}(.|\mathbf{X})}\left(\log\left(g_{\theta}\left(\mathbf{X}|\mathbf{Z}\right)\right)\right)\right]}_{\text{log-likelihood}} - \underbrace{\mathbb{E}_{f_{\mathbf{X}}}\left[D_{\mathrm{KL}}\left(g_{\phi}\left(.|\mathbf{X}\right)\|p\right)\right]}_{\text{regularisation}} =: \mathrm{ELBO}\left(\phi,\theta\right)$$

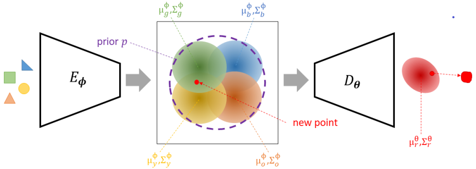
where *ELBO* refers to *Evidence Lower BOund*

Variational autoencoder

A new method for density approximation

New point generation procedure:

1. draw a point $\mathbf{z} \sim p$ from the prior p
2. draw a point $\mathbf{x} \sim g_{\boldsymbol{\theta}}(.|\mathbf{z})$

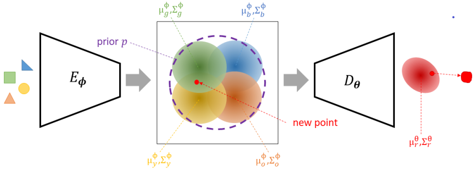


Variational autoencoder

A new method for density approximation

New point generation procedure:

1. draw a point $\mathbf{z} \sim p$ from the prior p
2. draw a point $\mathbf{x} \sim g_{\boldsymbol{\theta}}(.|\mathbf{z})$



As a result, a **variational autoencoder** returns a distribution on $\mathbb{R}^d$ of PDF:

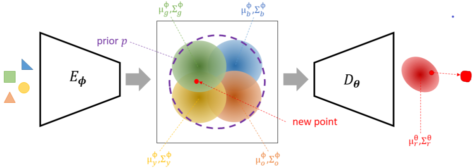
$$g_{\boldsymbol{\theta}}(\mathbf{x}) = \int g_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\, d\mathbf{z} = \int g_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$$

Variational autoencoder

A new method for density approximation

New point generation procedure:

1. draw a point $\mathbf{z} \sim p$ from the prior p
2. draw a point $\mathbf{x} \sim g_{\boldsymbol{\theta}}(.|\mathbf{z})$



As a result, a **variational autoencoder** returns a distribution on $\mathbb{R}^d$ of PDF:

$$g_{\boldsymbol{\theta}}(\mathbf{x}) = \int g_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\, d\mathbf{z} = \int g_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$$

Even if it is theoretically a parametric model parameterised by $\boldsymbol{\theta}$, it more looks like a non-parametric model and it is:

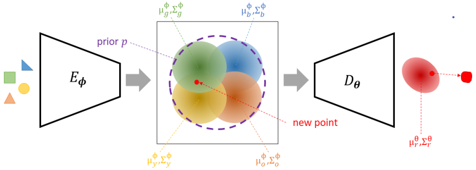
- ✔ flexible, since it is an infinite mixture of distributions $g_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$
- ✔ robust in high dimension, because of the dimensionality reduction

Variational autoencoder

A new method for density approximation

New point generation procedure:

1. draw a point $\mathbf{z} \sim p$ from the prior p
2. draw a point $\mathbf{x} \sim g_\theta(.|\mathbf{z})$



As a result, a **variational autoencoder** returns a distribution on $\mathbb{R}^d$ of PDF:

$$g_\theta(\mathbf{x}) = \int g_\theta(\mathbf{x}, \mathbf{z})\, d\mathbf{z} = \int g_\theta(\mathbf{x}|\mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$$

Question

Can we perform density estimation with a VAE in a context of importance sampling?

Outline

Density estimation with a VAE and weigthed samples

Mathematical details

Goal: Approximate a target distribution with a distribution parameterised by a VAE

IS case: Approximate g with data distributed according to $f_{\mathbf{X}}$ [DCBMK24]

Density estimation with a VAE and weigthed samples

Mathematical details

Goal: Approximate a target distribution with a distribution parameterised by a VAE

IS case: Approximate g with data distributed according to $f_{\mathbf{X}}$ [DCBMK24]

1. minimise $D_{\text{KL}}\left(g \| g_{\boldsymbol{\theta}}\right) = \mathbb{E}_g\left[\log\left(g\left(\mathbf{X}\right)\right) - \log\left(g_{\boldsymbol{\theta}}\left(\mathbf{X}\right)\right)\right]$ according to $\boldsymbol{\theta}$

Density estimation with a VAE and weigthed samples

Mathematical details

Goal: Approximate a target distribution with a distribution parameterised by a VAE

IS case: Approximate g with data distributed according to $f_{\mathbf{X}}$ [DCBMK24]

1. minimise $D_{\mathrm{KL}}(g \| g_{\boldsymbol{\theta}}) = \mathbb{E}_g[\log(g(\mathbf{X})) - \log(g_{\boldsymbol{\theta}}(\mathbf{X}))]$ according to $\boldsymbol{\theta}$
2. note that it is equivalent to maximise the log-likelihood $\mathbb{E}_g[\log(g_{\boldsymbol{\theta}}(\mathbf{X}))]$ according to $\boldsymbol{\theta}$

Density estimation with a VAE and weigthed samples

Mathematical details

Goal: Approximate a target distribution with a distribution parameterised by a VAE

IS case: Approximate g with data distributed according to $f_{\mathbf{X}}$ [DCBMK24]

1. minimise $D_{\mathrm{KL}}\left(g \| g_{\boldsymbol{\theta}}\right)=\mathbb{E}_{g}\left[\log \left(g\left(\mathbf{X}\right)\right)-\log \left(g_{\boldsymbol{\theta}}\left(\mathbf{X}\right)\right)\right]$ according to $\boldsymbol{\theta}$
2. note that it is equivalent to maximise the log-likelihood $\mathbb{E}_{g}\left[\log \left(g_{\boldsymbol{\theta}}\left(\mathbf{X}\right)\right)\right]$ according to $\boldsymbol{\theta}$
3. rewrite the log-likelihood as an expectation over $f_{\mathbf{X}}$ as $\mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})} \log \left(g_{\boldsymbol{\theta}}\left(\mathbf{X}\right)\right)\right]$ (IS trick)

Density estimation with a VAE and weigthed samples

Mathematical details

Goal: Approximate a target distribution with a distribution parameterised by a VAE

IS case: Approximate g with data distributed according to $f_{\mathbf{X}}$ [DCBMK24]

1. minimise $D_{\mathrm{KL}}(g \| g_{\theta}) = \mathbb{E}_g[\log(g(\mathbf{X})) - \log(g_{\theta}(\mathbf{X}))]$ according to θ
2. note that it is equivalent to maximise the log-likelihood $\mathbb{E}_g[\log(g_{\theta}(\mathbf{X}))]$ according to θ
3. rewrite the log-likelihood as an expectation over $f_{\mathbf{X}}$ as $\mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})}\log(g_{\theta}(\mathbf{X}))\right]$ (IS trick)
4. compute a lower bound of the weighted log-likelihood using the latent variable z:

$$\mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})}\log(g_{\theta}(\mathbf{X}))\right] \geq \underbrace{\mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})}\mathbb{E}_{g_{\phi}(.|\mathbf{X})}[\log(g_{\theta}(\mathbf{X}|\mathbf{Z}))]\right] - \mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})}D_{\mathrm{KL}}(g_{\phi}(.|\mathbf{X})\|p)\right]}_{\text{loss function of a VAE with weighted samples: wELBO}(\phi,\theta)}$$

Density estimation with a VAE and weigthed samples

Mathematical details

IS case: Approximate g with data distributed according to $f_{\mathbf{X}}$ [DCBMK24]

1. minimise $D_{\mathrm{KL}}(g \| g_{\theta}) = \mathbb{E}_g[\log(g(\mathbf{X})) - \log(g_{\theta}(\mathbf{X}))]$ according to $\boldsymbol{\theta}$
2. note that it is equivalent to maximise the log-likelihood $\mathbb{E}_g[\log(g_{\theta}(\mathbf{X}))]$ according to $\boldsymbol{\theta}$
3. rewrite the log-likelihood as an expectation over $f_{\mathbf{X}}$ as $\mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})} \log(g_{\theta}(\mathbf{X}))\right]$ (IS trick)
4. compute a lower bound of the weighted log-likelihood using the latent variable z:

$$\mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})} \log(g_{\theta}(\mathbf{X}))\right] \geq \underbrace{\mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})} \mathbb{E}_{g_{\phi}(.|\mathbf{X})}[\log(g_{\theta}(\mathbf{X}|\mathbf{Z}))]\right] - \mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})} D_{\mathrm{KL}}(g_{\phi}(.|\mathbf{X}) \| p)\right]}_{\text{loss function of a VAE with weighted samples: wELBO}(\boldsymbol{\phi},\boldsymbol{\theta})}$$

Statement

We can perform density estimation with weighted samples in an importance sampling context with a VAE by maximising $\mathrm{wELBO}(\boldsymbol{\phi}, \boldsymbol{\theta})$

Improvements of the VAE

Flexible prior and pre-training procedure

Challenge: ability to learn multimodal target distributions

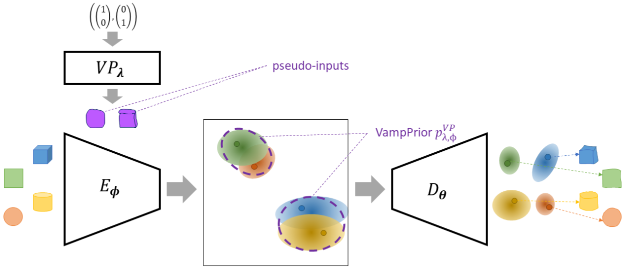
Improvements of the VAE

Flexible prior and pre-training procedure

Challenge: ability to learn multimodal target distributions

☞ Choice of a flexible prior: VampPrior [TW18]

To **add flexibility** to the resulting distribution $g_{\boldsymbol{\theta}}$, we consider a flexible prior distribution



$$p^{\mathrm{VP}}_{\boldsymbol{\lambda},\boldsymbol{\phi}}(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} g_{\boldsymbol{\phi}}\left(\mathbf{z}\,\middle|\,\mathrm{VP}_{\boldsymbol{\lambda}}\left(\mathbf{e}_k^K\right)\right)$$

- $\mathbf{e}_k^K$ are the vector of the canonical basis of $\mathbb{R}^K$
- $\mathrm{VP}_{\boldsymbol{\lambda}} : \mathbb{R}^K \to \mathbb{R}^d$ is a neural network

Improvements of the VAE

Flexible prior and pre-training procedure

Challenge: ability to learn multimodal target distributions

☞ Choice of a flexible prior: VampPrior [TW18]

To **add flexibility** to the resulting distribution $g_{\boldsymbol{\theta}}$, we consider a flexible prior distribution

- ✔ approximation the optimal prior p^*
- ✔ depends on $\boldsymbol{\phi} \Longrightarrow$ collaborative work between $E_{\boldsymbol{\phi}}$ and $\text{VP}_{\boldsymbol{\lambda}}$
- ✔ adapts itself to the data during the training $\Longrightarrow$ can be multimodal

Improvements of the VAE

Flexible prior and pre-training procedure

Challenge: ability to learn multimodal target distributions

☞ Choice of a flexible prior: VampPrior [TW18]

To **add flexibility** to the resulting distribution $g_{\boldsymbol{\theta}}$, we consider a flexible prior distribution

- ✔ approximation the optimal prior p^*
- ✔ depends on $\boldsymbol{\phi} \implies$ collaborative work between $E_{\boldsymbol{\phi}}$ and $\mathrm{VP}_{\boldsymbol{\lambda}}$
- ✔ adapts itself to the data during the training $\implies$ can be multimodal

Introduction of $p^{\mathrm{VP}}_{\boldsymbol{\lambda},\boldsymbol{\phi}}$ into the loss function:

$$\arg\max_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\lambda}} \underbrace{\mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})}\mathbb{E}_{g_{\boldsymbol{\phi}}(.|\mathbf{X})}\left(\log\left(g_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})\right)\right)\right]}_{\text{log-likelihood}} - \underbrace{\mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})}D_{\mathrm{KL}}\left(g_{\boldsymbol{\phi}}(.|\mathbf{X})\,\|\,p^{\mathrm{VP}}_{\boldsymbol{\lambda},\boldsymbol{\phi}}\right)\right]}_{\text{regularisation term}}$$

Improvements of the VAE

Flexible prior and pre-training procedure

Challenge: ability to learn multimodal target distributions

☞ Pre-training procedure [DCBMK24]

The **posterior collapse** phenomenon can badly affect the performances of the VAE

- ✘ over-regularisation of the VAE, bad reconstruction of the data
- ✘ unimodal resulting distribution
- ✘ stuck in a local optimum during the training of the VAE

Improvements of the VAE

Flexible prior and pre-training procedure

Challenge: ability to learn multimodal target distributions

☞ Pre-training procedure [DCBMK24]

The **posterior collapse** phenomenon can badly affect the performances of the VAE

- ✘ over-regularisation of the VAE, bad reconstruction of the data
- ✘ unimodal resulting distribution
- ✘ stuck in a local optimum during the training of the VAE

Our remedy: new pre-training procedure to find "good" starting points $\boldsymbol{\phi}^{(0)}$,$\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$

1. initialise the weights $\boldsymbol{\lambda}$ by supervised learning
2. initialise the weights $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ by unsupervised learning
3. main training of the VAE

Importance sampling with a VAE

Compute the PDF values of the resulting distribution

Question: How can we have access to the PDF values of $g_{\boldsymbol{\theta}}(\mathbf{x}) = \int g_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$?

Importance sampling with a VAE

Compute the PDF values of the resulting distribution

Question: How can we have access to the PDF values of $g_{\boldsymbol{\theta}}(\mathbf{x}) = \int g_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$?

☞ Existing procedure [WBD19]: pointwise estimation $\widehat{g_{\boldsymbol{\theta}}(\mathbf{x})}$ of the PDF values of $g_{\boldsymbol{\theta}}$

- ✘ the convenient statistical properties of $\widehat{p}_{t,N}^{\mathrm{IS}}$, unbiasedness and convergence, are no longer guaranteed

Importance sampling with a VAE

Compute the PDF values of the resulting distribution

Question: How can we have access to the PDF values of $g_{\boldsymbol{\theta}}(\mathbf{x}) = \int g_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$?

☞ Existing procedure [WBD19]: pointwise estimation $\widehat{g_{\boldsymbol{\theta}}(\mathbf{x})}$ of the PDF values of $g_{\boldsymbol{\theta}}$

✘ the convenient statistical properties of $\widehat{p}^{\text{IS}}_{t,N}$, unbiasedness and convergence, are no longer guaranteed

☞ Our procedure [DCBMK24]: we propose no longer to estimate only the PDF values of $g_{\boldsymbol{\theta}}$ pointwise, but to **approximate the whole distribution** $g_{\boldsymbol{\theta}}$ by the mixture:

$$g_{\boldsymbol{\theta}}^M(.) = \frac{1}{M}\sum_{m=1}^{M} g_{\boldsymbol{\theta}}\left(. \,\middle|\, \mathbf{Z}^{(m)}\right) \text{ with } \left(\mathbf{Z}^{(m)}\right)_{m\in[\![1,M]\!]} \in \mathcal{Z}^M \sim p \quad \text{i.i.d.}$$

✔ It is possible to compute exactly the PDF values of $g_{\boldsymbol{\theta}}^M$!

Importance sampling with a VAE

Methodology

IS goal: approximate $g_{\text{opt}}(\mathbf{x}) \propto \mathbb{1}(\psi(\mathbf{x}) > t)\, f_{\mathbf{X}}(\mathbf{x})$ with data distributed according to $f_{\mathbf{X}}$

Importance sampling with a VAE

Methodology

IS goal: approximate $g_{\text{opt}}(\mathbf{x}) \propto \mathbb{1}(\psi(\mathbf{x}) > t) f_{\mathbf{X}}(\mathbf{x})$ with data distributed according to $f_{\mathbf{X}}$

Methodology [DCBMK24]:

1. train a VAE by maximising

$$\text{wELBO}(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbb{E}_{f_{\mathbf{X}}}\left[\mathbb{1}(\phi(\mathbf{X}) > t)\, \mathbb{E}_{g_{\boldsymbol{\phi}}(.|\mathbf{X})}[\log(g_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z}))]\right] - \mathbb{E}_{f_{\mathbf{X}}}\left[\mathbb{1}(\phi(\mathbf{X}) > t)\, D_{\text{KL}}\left(g_{\boldsymbol{\phi}}(.|\mathbf{X}) \| p^{\text{VP}}_{\boldsymbol{\lambda},\boldsymbol{\phi}}\right)\right]$$

2. compute the resulting approximating distribution $g^M_{\boldsymbol{\theta}}$
3. draw a N-sample according to $g^M_{\boldsymbol{\theta}}$
4. estimate the failure probability with the importance sampling estimator $\widehat{p}^{\text{IS}}_{t,N}$

Importance sampling with a VAE

Methodology

IS goal: approximate $g_{\text{opt}}(\mathbf{x}) \propto \mathbb{1}(\psi(\mathbf{x}) > t) f_{\mathbf{X}}(\mathbf{x})$ with data distributed according to $f_{\mathbf{X}}$

Methodology [DCBMK24]:

1. train a VAE by maximising

$$\text{wELBO}(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbb{E}_{f_{\mathbf{X}}}\left[\mathbb{1}(\phi(\mathbf{X}) > t)\, \mathbb{E}_{g_{\boldsymbol{\phi}}(.|\mathbf{X})}\left[\log\left(g_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})\right)\right]\right] - \mathbb{E}_{f_{\mathbf{X}}}\left[\mathbb{1}(\phi(\mathbf{X}) > t)\, D_{\text{KL}}\left(g_{\boldsymbol{\phi}}(.|\mathbf{X}) \| p^{\text{VP}}_{\boldsymbol{\lambda},\boldsymbol{\phi}}\right)\right]$$

2. compute the resulting approximating distribution $g^M_{\boldsymbol{\theta}}$
3. draw a N-sample according to $g^M_{\boldsymbol{\theta}}$
4. estimate the failure probability with the importance sampling estimator $\widehat{p}^{\text{IS}}_{t,N}$

Theorem ([DCBMK24])

The estimator $\widehat{p}^{\text{IS}}_{t,N}$ with $g^M_{\boldsymbol{\theta}}$ as the auxiliary distribution is unbiased and convergent

Outline

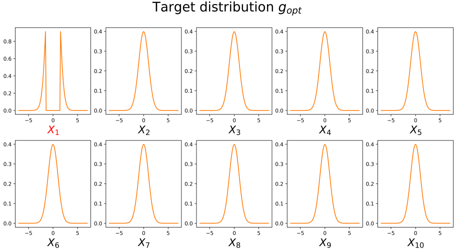
Numerical test

Estimation of the failure probability on a simple test case in dimension 10

Problem setting:

- Black-box model: $\forall \mathbf{x} \in \mathbb{R}^{10},\ \psi(\mathbf{x}) = |x_1|$
- failure threshold: $t = 1.5$
- input distribution: $f_{\mathbf{X}} = \mathcal{N}_{10}(\mathbf{0}_{10}, \mathbf{I}_{10})$

$\hookrightarrow p_t \approx 1.336 \times 10^{-1}$



Numerical test

Estimation of the failure probability on a simple test case in dimension 10

Problem setting:

- Black-box model: $\forall \mathbf{x} \in \mathbb{R}^{10},\ \psi(\mathbf{x}) = |x_1|$
- failure threshold: $t = 1.5$
- input distribution: $f_{\mathbf{X}} = \mathcal{N}_{10}(\mathbf{0}_{10}, \mathbf{I}_{10})$

$\hookrightarrow p_t \approx 1.336 \times 10^{-1}$

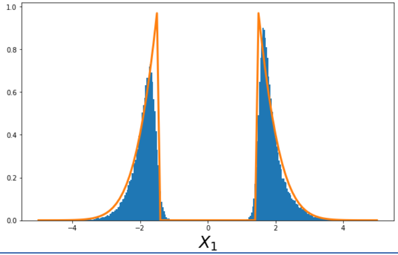
Estimation of the failure probability:

$\widehat{p}_{t,N}^{\mathrm{IS}}$	C.o.V. $\left(\widehat{p}_{t,N}^{\mathrm{IS}}\right)$
1.339×10^{-1}	0.540%

Table: Theoretical error Monte Carlo: C.o.V. $\left(\widehat{p}_{t,N}^{\mathrm{MC}}\right) = 2.546\%$

Parameters of the algorithm:

- dimension of the latent space: $d_z = 2$
- VampPrior components: $K = 75$
- $N = 10^4$
- $M = 10^3$



Numerical test

Estimation of the failure probability on a simple test case in dimension 100

Problem setting:

- Black-box model: $\forall \mathbf{x} \in \mathbb{R}^{100}, \ \psi(\mathbf{x}) = |x_1|$
- failure threshold: $t = 1.5$
- input distribution: $f_{\mathbf{X}} = \mathcal{N}_{100}(\mathbf{0}_{100}, \mathbf{I}_{100})$

$\hookrightarrow p_t \approx 1.336 \times 10^{-1}$

Parameters of the algorithm:

- dimension of the latent space: $d_z = 2$
- VampPrior components: $K = 75$
- $N = 10^4$
- $M = 10^3$

Numerical test

Estimation of the failure probability on a simple test case in dimension 100

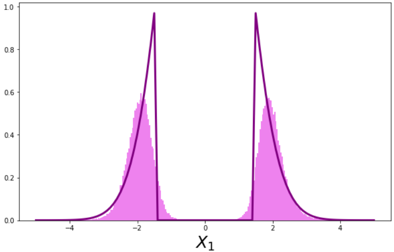
Problem setting:

- Black-box model: $\forall \mathbf{x} \in \mathbb{R}^{100}, \ \psi(\mathbf{x}) = |x_1|$
- failure threshold: $t = 1.5$
- input distribution: $f_{\mathbf{X}} = \mathcal{N}_{100}(\mathbf{0}_{100}, \mathbf{I}_{100})$

$\hookrightarrow p_t \approx 1.336 \times 10^{-1}$

Parameters of the algorithm:

- dimension of the latent space: $d_z = 2$
- VampPrior components: $K = 75$
- $N = 10^4$
- $M = 10^3$



Estimation of the failure probability:

$\widehat{p}_{t,N}^{\mathrm{IS}}$	C.o.V. $\left(\widehat{p}_{t,N}^{\mathrm{IS}}\right)$
1.355×10^{-1}	1.486%

Table: Theoretical error Monte Carlo: C.o.V. $\left(\widehat{p}_{t,N}^{\mathrm{MC}}\right) = 2.546\%$

Adaptive IS with a VAE

How to estimate a rare event probability with a VAE?

- ✔ The VAE found both modes in dimension 10 and 100, and the estimation error is small
- ✘ Fine...... but $p_t \approx 1.336 \times 10^{-1}$ is not the probability of a rare event

Adaptive IS with a VAE

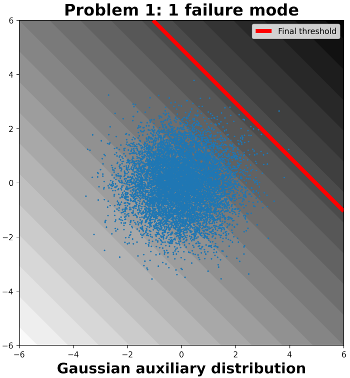
How to estimate a rare event probability with a VAE?

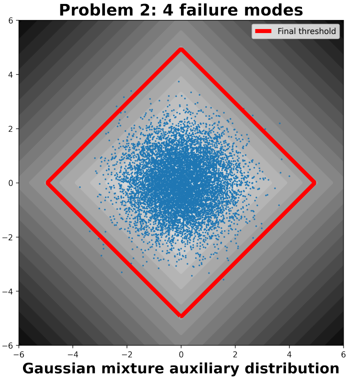
Question: how to deal with rare event probabilities?

Adaptive IS with a VAE

How to estimate a rare event probability with a VAE?

Question: how to deal with rare event probabilities?
Solution: use an adaptive IS algorithm $\Longrightarrow$ the cross-entropy algorithm [RK04]

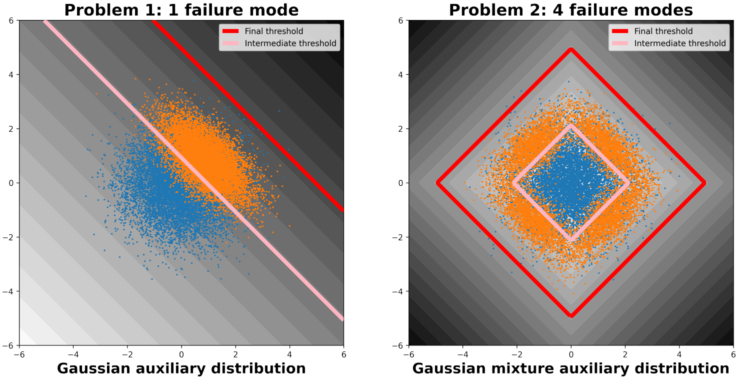




Adaptive IS with a VAE

How to estimate a rare event probability with a VAE?

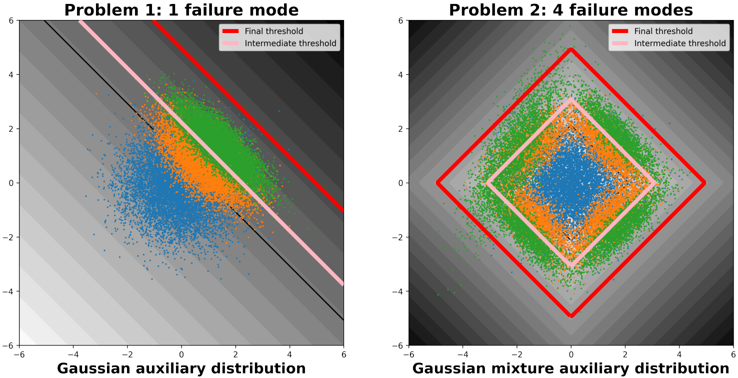
Question: how to deal with rare event probabilities?
Solution: use an adaptive IS algorithm $\Longrightarrow$ the cross-entropy algorithm [RK04]



Adaptive IS with a VAE

How to estimate a rare event probability with a VAE?

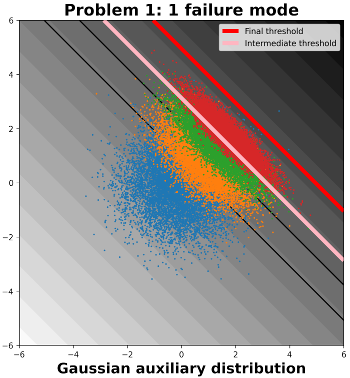
Question: how to deal with rare event probabilities?
Solution: use an adaptive IS algorithm $\Longrightarrow$ the cross-entropy algorithm [RK04]

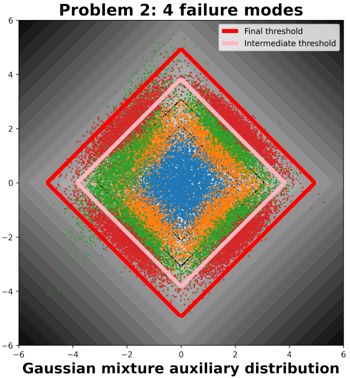


Adaptive IS with a VAE

How to estimate a rare event probability with a VAE?

Question: how to deal with rare event probabilities?
Solution: use an adaptive IS algorithm $\Longrightarrow$ the cross-entropy algorithm [RK04]

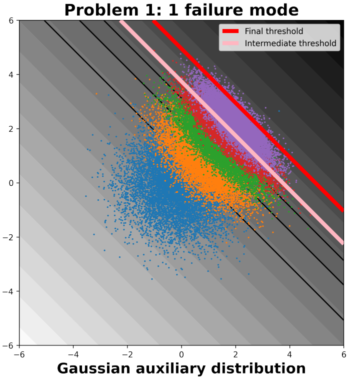


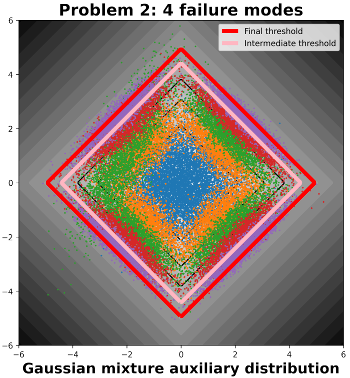


Adaptive IS with a VAE

How to estimate a rare event probability with a VAE?

Question: how to deal with rare event probabilities?
Solution: use an adaptive IS algorithm $\Longrightarrow$ the cross-entropy algorithm [RK04]

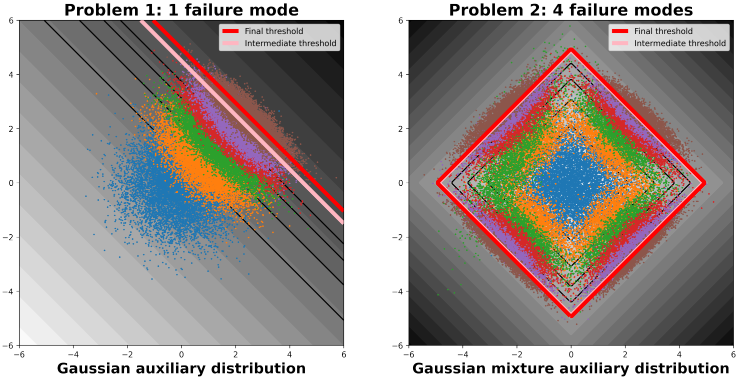




Adaptive IS with a VAE

How to estimate a rare event probability with a VAE?

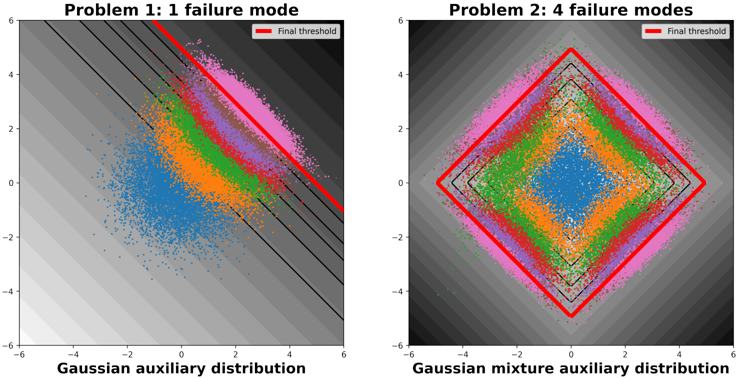
Question: how to deal with rare event probabilities?
Solution: use an adaptive IS algorithm $\Longrightarrow$ the cross-entropy algorithm [RK04]



Adaptive IS with a VAE

How to estimate a rare event probability with a VAE?

Question: how to deal with rare event probabilities?
Solution: use an adaptive IS algorithm $\Longrightarrow$ the cross-entropy algorithm [RK04]



Adaptive IS with a VAE

How to estimate a rare event probability with a VAE?

Question: how to deal with rare event probabilities?
Solution: use an adaptive IS algorithm $\Longrightarrow$ the cross-entropy algorithm [RK04]

☞ Existing CE algorithms can use as the auxiliary distribution:

- Gaussian distributions
- Gaussian mixture distributions
- non-parametric models
- Mixture of von Mises-Fisher-Nakagami (vMFNM) distributions

Adaptive IS with a VAE

How to estimate a rare event probability with a VAE?

Question: how to deal with rare event probabilities?
Solution: use an adaptive IS algorithm $\Longrightarrow$ the cross-entropy algorithm [RK04]

☞ Existing CE algorithms can use as the auxiliary distribution:

- Gaussian distributions
- Gaussian mixture distributions
- non-parametric models
- Mixture of von Mises-Fisher-Nakagami (vMFNM) distributions

☞ Our improvement: **CE-VAE** algorithm
New CE algorithm using a distribution parameterised by a VAE as the auxiliary distribution

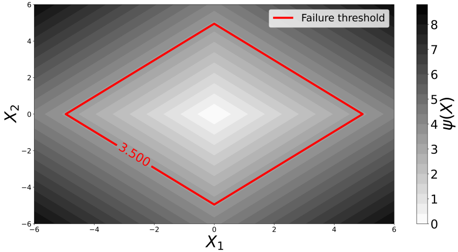
Numerical test

4-branch problem in dimension 100

Problem setting:

- "4-branch" in dimension $d = 100$
- failure threshold: $t = 3.5$
- input distribution: $f_{\mathbf{X}} = \mathcal{N}_{100}\left(\mathbf{0}_{100}, \mathbf{I}_{100}\right)$

$\hookrightarrow p_t \approx 9.3 \times 10^{-4}$



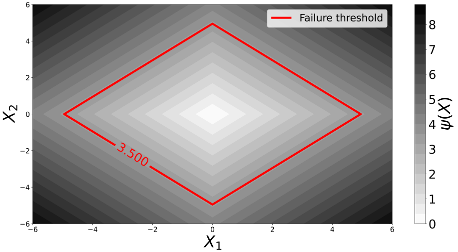
Numerical test

4-branch problem in dimension 100

Problem setting:

- "4-branch" in dimension $d = 100$
- failure threshold: $t = 3.5$
- input distribution: $f_{\mathbf{X}} = \mathcal{N}_{100}\left(\mathbf{0}_{100}, \mathbf{I}_{100}\right)$

$\hookrightarrow p_t \approx 9.3 \times 10^{-4}$



Comparison with the CE algorithm using as the auxiliary distribution:

- a mixture of $n \in \{3, 4, 5\}$ vMFNM distributions (CE-vMFNMn) [PGS19]
- a standard VAE without both VampPrior and the pre-training procedure (CE-stdVAE)

Numerical test

4-branch problem in dimension 100

	CE-VAE	CE-vMFNM3	CE-vMFNM4	CE-vMFNM5	CE-stdVAE
N_{tot}	**40000**	88000	50000	50000	200000
$\widehat{p}_t^{\text{mean}}$	9.310×10^{-4}	1.319×10^{-3}	9.835×10^{-4}	9.315×10^{-4}	9.446×10^{-4}
C.o.V. $(\widehat{p}_t)$	5.31%	512.8%	31.3%	7.56%	34.83%

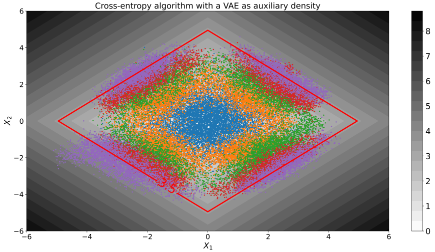
The CE-VAE algorithm:

- ✔ requires less iterations to converge
- ✔ has the smallest estimation error
- ✔ doesn't require any prior knowledge on the form of the failure domain
- ✔ major beneficial impact of both VampPrior and the pre-training procedure

Numerical test

4-branch problem in dimension 100

	CE-VAE	CE-vMFNM3	CE-vMFNM4	CE-vMFNM5	CE-stdVAE
N_{tot}	**40000**	88000	50000	50000	200000
$\widehat{p}_t^{mean}$	9.310×10^{-4}	1.319×10^{-3}	9.835×10^{-4}	9.315×10^{-4}	9.446×10^{-4}
C.o.V. $(\widehat{p}_t)$	**5.31%**	512.8%	31.3%	7.56%	34.83%



Outline

Conclusion and perspectives

What is new?

- ✔ adaptation of the VAE framework to approximate a target distribution with weighted samples
- ✔ able to learn a multimodal target distribution without any prior knowledge on it
- ✔ procedure can be applied to any kind of importance sampling (reliability analysis, generation)

Conclusion and perspectives

What is new?

- ✔ adaptation of the VAE framework to approximate a target distribution with weighted samples
- ✔ able to learn a multimodal target distribution without any prior knowledge on it
- ✔ procedure can be applied to any kind of importance sampling (reliability analysis, generation)

Improvements and perspectives:

- apply numerical tricks to prevent the weight degeneracy phenomenon in very high dimension
- improve the ability of the method to learn multimodal target distributions, in particular in a non-reliability context
- extend the procedure to the estimation of reliability-oriented sensitivity indices based on [PD19] or on [DCBM23]

✌ Published paper [DCBMK24] and codes to reproduce the results are available online!

References

[Bre84] Karl Breitung.
Asymptotic approximations for multinormal integrals.
Journal of Engineering Mechanics, 110(3):357–366, 1984.

[Buc04] James Bucklew.
Introduction to rare event simulation.
Springer Science & Business Media, 2004.

[BVV+15] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio.
Generating sentences from a continuous space.
arXiv preprint arXiv:1511.06349, 2015.

[CDMFG12] Frédéric Cérou, Pierre Del Moral, Teddy Furon, and Arnaud Guyader.
Sequential monte carlo for rare event estimation.
Statistics and computing, 22(3):795–808, 2012.

[DCBM23] Julien Demange-Chryst, François Bachoc, and Jérôme Morio. Shapley effect estimation in reliability-oriented sensitivity analysis with correlated inputs by importance sampling. International Journal for Uncertainty Quantification, 13(3), 2023.

[DCBMK24] Julien Demange-Chryst, Francois Bachoc, Jérôme Morio, and Timothé Krauth. Variational autoencoder with weighted samples for high-dimensional non-parametric adaptive importance sampling. Transactions on Machine Learning Research, 2024.

[FCIM23] Elias Fekhari, Vincent Chabridon, Bertrand Iooss, and Joseph Muré. Bernstein adaptive nonparametric conditional sampling: a new method for rare event probability estimation. In International Conference on Application of Statistics and Probability in Civil Engineering, 2023.

[GPS19] Sebastian Geyer, Iason Papaioannou, and Daniel Straub.
Cross entropy-based importance sampling using Gaussian densities revisited.
Structural Safety, 76:15–27, 2019.

[HJ16] Matthew D Hoffman and Matthew J Johnson.
ELBO surgery: yet another way to carve up the variational evidence lower bound.
In Workshop in Advances in Approximate Bayesian Inference, NIPS, volume 1, 2016.

[HL74] Abraham M Hasofer and Niels C Lind.
Exact and invariant second-moment code format.
Journal of the Engineering Mechanics division, 100(1):111–121, 1974.

[KW14] Diederik P. Kingma and Max Welling.
Auto-Encoding Variational Bayes.
Article accepted in the 2nd International Conference on Learning Representations 2014, 2014.

[Mor11] Jérôme Morio.
Non-parametric adaptive importance sampling for the probability estimation of a launcher impact position.
Reliability engineering & system safety, 96(1):178–183, 2011.

[MRG+87] James L McClelland, David E Rumelhart, PDP Research Group, et al.
Parallel distributed processing, volume 2: Explorations in the microstructure of cogn
volume 2.
MIT press, 1987.

[MSJ+15] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey.
Adversarial autoencoders.
arXiv preprint arXiv:1511.05644, 2015.

[PD19] Guillaume Perrin and G Defaux.
Efficient evaluation of reliability-oriented sensitivity indices.
Journal of Scientific Computing, 79(3):1433–1455, 2019.

[PGS19] Iason Papaioannou, Sebastian Geyer, and Daniel Straub. Improved cross entropy-based importance sampling with a flexible mixture model. Reliability Engineering & System Safety, 191:106564, 2019.

[RK04] Reuven Y Rubinstein and Dirk P Kroese. The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning, volume 133. Springer, 2004.

[SRM+16] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. Advances in neural information processing systems, 29, 2016.

[TW18] Jakub Tomczak and Max Welling. VAE with a VampPrior. In International Conference on Artificial Intelligence and Statistics, pages 1214–1223. PMLR, 2018.

[WBD19] Hechuan Wang, Mónica F Bugallo, and Petar M Djurić.
Adaptive importance sampling supported by a variational auto-encoder.
In 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pages 619–623. IEEE, 2019.

[WEG87] Svante Wold, Kim Esbensen, and Paul Geladi.
Principal component analysis.
Chemometrics and intelligent laboratory systems, 2(1-3):37–52, 1987.

[Zha96] Ping Zhang.
Nonparametric importance sampling.
Journal of the American Statistical Association, 91(435):1245–1253, 1996.

Choice of the prior

Optimal prior distribution

The most classical and easiest choice for the prior is $p = \mathcal{N}_{d_z}\left(\mathbf{0}_{d_z}, \boldsymbol{I}_{d_z}\right)$

- ✘ can be too restrictive, for multimodal target distributions for example, and can lead to over-regularisation and finally to poor density estimation
- **question**: how can we add flexibility to $g_{\boldsymbol{\theta}}$ with the prior?

Choice of the prior

Optimal prior distribution

The most classical and easiest choice for the prior is $p = \mathcal{N}_{d_z}\left(\mathbf{0}_{d_z}, \boldsymbol{I}_{d_z}\right)$

- ✘ can be too restrictive, for multimodal target distributions for example, and can lead to over-regularisation and finally to poor density estimation
- **question**: how can we add flexibility to $g_{\boldsymbol{\theta}}$ with the prior?

Solution

Consider a flexible and learnable prior p. The optimal prior distribution, maximising the loss function, is given by [MSJ+15, HJ16]:

$$p^*(\mathbf{z}) = \int g_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\, g_{\text{opt}}(\mathbf{x})\, d\mathbf{x} = \mathbb{E}_{g_{\text{opt}}}\left[g_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})\right]$$

This is the *aggregated posterior*.

Choice of the prior

VampPrior

There are several existing methods to approximate this optimal prior

Chosen method: *Variational Mixture of Posteriors* prior, or *VampPrior* [TW18]

$$p^{\mathrm{VP}}_{\mathbf{u}_1,\dots,\mathbf{u}_K,\boldsymbol{\phi}}(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} g_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{u}_k)$$

where $K \geq 1$ and $(\mathbf{u}_k)_{k\in[\![1,K]\!]}$ are learnable pseudo-inputs from the initial space $\mathbb{R}^d$

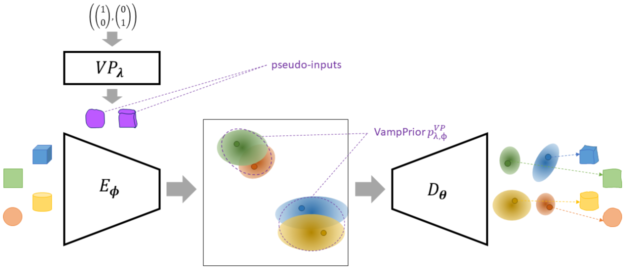
Advantages of the VampPrior distribution:

- ✔ flexible enough to be adapted to many kinds of problems
- ✔ depends on $\boldsymbol{\phi}$, as the aggregated posterior p^*

Choice of the prior

VampPrior

Chosen method: *Variational Mixture of Posteriors* prior, or *VampPrior* [TW18]



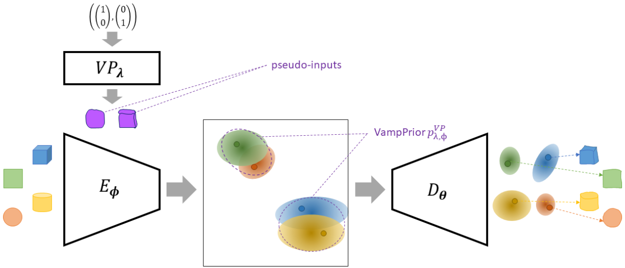
$$p^{\mathrm{VP}}_{\boldsymbol{\lambda},\boldsymbol{\phi}}(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} g_{\boldsymbol{\phi}}\left(\mathbf{z} \,\middle|\, \mathrm{VP}_{\boldsymbol{\lambda}}\left(\mathbf{e}_k^K\right)\right)$$

- $\mathbf{e}_k^K$ are the vector of the canonical basis of $\mathbb{R}^K$
- $\mathrm{VP}_{\boldsymbol{\lambda}} : \mathbb{R}^K \to \mathbb{R}^d$ is a neural network

Choice of the prior

VampPrior

Chosen method: *Variational Mixture of Posteriors* prior, or *VampPrior* [TW18]



$$p^{\mathrm{VP}}_{\boldsymbol{\lambda},\boldsymbol{\phi}}(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} g_{\boldsymbol{\phi}}\left(\mathbf{z}\,\middle|\,\mathrm{VP}_{\boldsymbol{\lambda}}\left(\boldsymbol{e}^K_k\right)\right)$$

- $\boldsymbol{e}^K_k$ are the vector of the canonical basis of $\mathbb{R}^K$
- $\mathrm{VP}_{\boldsymbol{\lambda}} : \mathbb{R}^K \to \mathbb{R}^d$ is a neural network

Introduction $p^{\mathrm{VP}}_{\boldsymbol{\lambda},\boldsymbol{\phi}}$ into the loss function:

$$\underset{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\lambda}}{\arg\max}\ \mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})}\mathbb{E}_{g_{\boldsymbol{\phi}}(.|\mathbf{X})}\left(\log\left(g_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})\right)\right)\right] - \mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})}D_{\mathrm{KL}}\left(g_{\boldsymbol{\phi}}(.|\mathbf{X})\,\|\,p^{\mathrm{VP}}_{\boldsymbol{\lambda},\boldsymbol{\phi}}\right)\right]$$

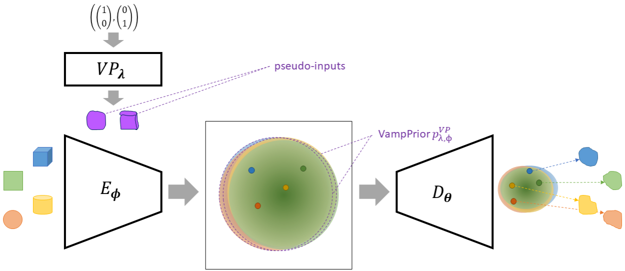
Posterior collapse
New pre-training procedure

Posterior collapse [BVV+15, SRM+16] is a phenomenon that badly affects the performances of a VAE

Posterior collapse

New pre-training procedure

Posterior collapse [BVV+15, SRM+16] is a phenomenon that badly affects the performances of a VAE

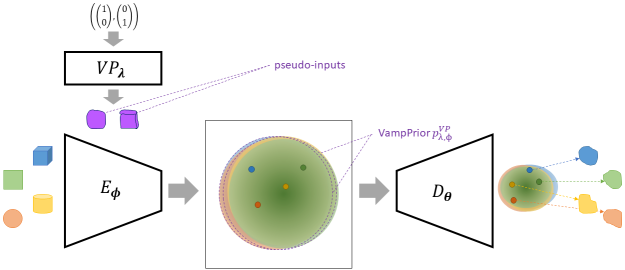


It generally refers to:

- an over-regularisation of the VAE
- i.e. $D_{KL}\left(g_\phi(.\,|\mathbf{x})\,\|\,p^{VP}_{\boldsymbol{\lambda},\boldsymbol{\phi}}\right) \approx 0$ for every $\mathbf{x} \in \mathbb{R}^d$

Posterior collapse
New pre-training procedure

Posterior collapse [BVV+15, SRM+16] is a phenomenon that badly affects the performances of a VAE



It generally refers to:

- an over-regularisation of the VAE
- i.e. $D_{\mathrm{KL}}\left(g_\phi(.\,|\mathbf{x})\,\|\,p^{\mathrm{VP}}_{\boldsymbol{\lambda},\boldsymbol{\phi}}\right) \approx 0$ for every $\mathbf{x} \in \mathbb{R}^d$

Why? Not a clear answer!
The most common hypothesis is that posterior collapse happens when we are stuck in a local maxima during the training of the VAE [SRM+16]

Posterior collapse

New pre-training procedure

Existing remedies are based on some modifications of the loss function or on the choice of other families for the prior p and/or the posterior distributions $g_{\boldsymbol{\phi}}(.\,|\mathbf{x})$

Posterior collapse

New pre-training procedure

Our remedy: new pre-training procedure to find "good" starting points $\boldsymbol{\phi}^{(0)}$,$\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$

Posterior collapse

New pre-training procedure

Our remedy: new pre-training procedure to find "good" starting points $\boldsymbol{\phi}^{(0)}$,$\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$

1. initialise the weights $\boldsymbol{\lambda}$ by supervised learning
2. initialise the weights $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ by unsupervised learning
3. main training of the VAE

Posterior collapse

New pre-training procedure

Our remedy: new pre-training procedure to find "good" starting points $\boldsymbol{\phi}^{(0)}$,$\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$

1. initialise the weights $\boldsymbol{\lambda}$ by supervised learning
   - pick a sub-sample $\left(\mathbf{X}^{(s(k))}\right)_{k\in[\![1,K]\!]}$ with probabilities $\propto \left(g_{\text{opt}}\left(\mathbf{X}^{(n)}\right) / f_{\mathbf{X}}\left(\mathbf{X}^{(n)}\right)\right)_{n\in[\![1,N]\!]}$
   - pre-train the $\text{VP}_{\boldsymbol{\lambda}}$ network by solving

$$\boldsymbol{\lambda}^{(0)} = \arg\min_{\boldsymbol{\lambda}} \sum_{k=1}^{K} \left\|\text{VP}_{\boldsymbol{\lambda}}\left(\mathbf{e}_k^K\right) - \mathbf{X}^{(s(k))}\right\|^2$$

   - the initial pseuso-inputs $\boldsymbol{u}_k^{(0)} = \text{VP}_{\boldsymbol{\lambda}^{(0)}}\left(\mathbf{e}_k^K\right)$ are already representative of the target distribution g_{opt}
2. initialise the weights $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ by unsupervised learning
3. main training of the VAE

Posterior collapse

New pre-training procedure

Our remedy: new pre-training procedure to find "good" starting points $\boldsymbol{\phi}^{(0)}$,$\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$

1. initialise the weights $\boldsymbol{\lambda}$ by supervised learning
2. initialise the weights $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ by unsupervised learning
   - pre-train the pair encoder/decoder $(E_{\boldsymbol{\phi}}, D_{\boldsymbol{\theta}})$ as a classical autoencoder by solving:

$$\boldsymbol{\phi}^{(0)}, \boldsymbol{\theta}^{(0)} = \arg\min_{\boldsymbol{\phi},\boldsymbol{\theta}} \mathbb{E}_{f_{\mathbf{X}}} \left[\frac{g_{\mathrm{opt}}(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})} \left\| \mathbf{X} - \boldsymbol{\mu}^{\boldsymbol{\theta}}_{\boldsymbol{\mu}^{\boldsymbol{\phi}}_{\mathbf{x}}} \right\|^2 \right]$$

   where $\left(\boldsymbol{\mu}^{\boldsymbol{\phi}}_{\mathbf{x}}, \boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\mathbf{x}}\right) = E_{\boldsymbol{\phi}}(\mathbf{x})$ and $\left(\boldsymbol{\mu}^{\boldsymbol{\theta}}_{\mathbf{z}}, \boldsymbol{\Sigma}^{\boldsymbol{\theta}}_{\mathbf{z}}\right) = D_{\boldsymbol{\theta}}(\mathbf{z})$ when respectively $g_{\boldsymbol{\phi}}(.\,|\mathbf{x})$ and $g_{\boldsymbol{\theta}}(.\,|\mathbf{z})$ are Gaussian distribution with diagonal covariance matrices.
3. main training of the VAE

Posterior collapse

New pre-training procedure

Our remedy: new pre-training procedure to find "good" starting points $\boldsymbol{\phi}^{(0)}$,$\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$

1. initialise the weights $\boldsymbol{\lambda}$ by supervised learning
2. initialise the weights $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ by unsupervised learning
3. main training of the VAE
   - train the whole VAE $(E_{\boldsymbol{\phi}}, D_{\boldsymbol{\theta}}, \mathrm{VP}_{\boldsymbol{\lambda}})$ by solving:

$$\boldsymbol{\phi}^*, \boldsymbol{\theta}^*, \boldsymbol{\lambda}^* = \underset{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\lambda}}{\arg\max}\ \mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})}\mathbb{E}_{g_{\boldsymbol{\phi}}(.|\mathbf{X})}\left(\log\left(g_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})\right)\right)\right] - \mathbb{E}_{f_{\mathbf{X}}}\left[\frac{g(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})}D_{\mathrm{KL}}\left(g_{\boldsymbol{\phi}}(.|\mathbf{X})\,\|p^{\mathrm{VP}}_{\boldsymbol{\lambda},\boldsymbol{\phi}}\right)\right]$$

starting from $\left(\boldsymbol{\phi}^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{\lambda}^{(0)}\right)$