Hyperspherical Variational Auto-Encoder

Géraud Ilinca, Louis Martinez

VAE framework

Observed variable	X
Latent variable	Z
Distribution to sample	$p(x) = \int p_\phi(x, z)dz = \int p_\phi(x \mid z)p(z)dz$
Likelihood (decoder)	$p_\phi(x \mid z)$
Posterior (encoder)	$q_\psi(z \mid x; \theta)$ ←usually $\mathcal{N}(\mu, \Sigma)$
Prior	$p(z)$ ←usually $\mathcal{N}(0, I)$

ELBO

Motivation: log-likelihood cannot be directly maximized

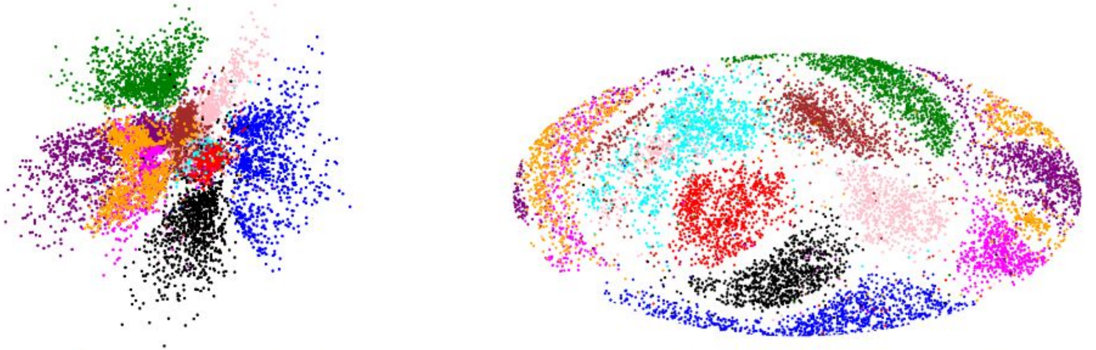
$$\log \int p_\phi(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$ intractable ⇒ variational approach

$$\log \int p_\phi(\mathbf{x}, \mathbf{z}) d\mathbf{z} \;\geq\; \text{ELBO} = \underbrace{\mathbb{E}_{q_\psi(z|x;\theta)}[log\, p_\phi(x \mid z)] - \text{KL}\left(q_\psi(z \mid x;\theta) \,||\, p(z)\right)}_{\text{To maximize}}$$

$$KL_{\mathcal{N}-VAE} = \frac{1}{2} \sum_{i=1}^{d} (1 + \log(\sigma^2_{\psi,j}) - \mu^2_{\psi,j} - \sigma^2_{\psi,j})$$

Motivations

- Limitation of Gaussian prior
  ↳Unsuitable to model latent hyperspherical structure
- Directional data modeling (wind direction, protein structure, etc.)
- Improved clusterability



(a) $\mathbb{R}^2$ latent space of the $\mathcal{N}$-VAE.

(b) Hammer projection of $\mathcal{S}^2$ latent space of the $\mathcal{S}$-VAE.

Image from the paper

vMF sampling

Key properties enabling Ulrich-Woods sampling :

- Action of orthogonal transformations :

$$\mathbf{x} \sim \mathrm{VMF}(\boldsymbol{\mu}, \kappa)$$
$$\mathbf{y} \sim \mathrm{VMF}(\mathbf{U}\boldsymbol{\mu}, \kappa)$$
$$\Rightarrow \mathbf{y} = \mathbf{U}\mathbf{x}$$

- Known marginal density :

$$g(\omega|\kappa, m) \propto \exp(\kappa\omega)(1-\omega^2)^{(m-3)/2}$$

$\Rightarrow$ Rejection Sampling

- Uniform on subsphere in hyperplane

orthogonal to $\boldsymbol{\mu}$

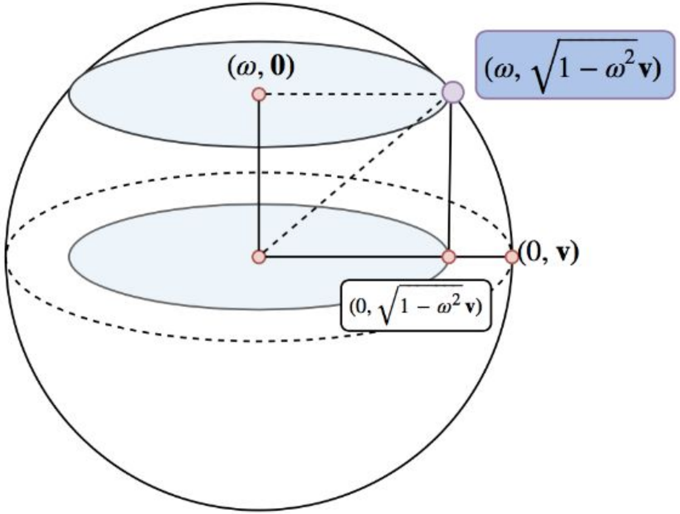


Image from the paper

Reparametrization Trick

N-VAE situation :

$$\mathrm{ELBO} = \mathbb{E}_{q_\psi(z|x;\theta)}[log\, p_\phi(x \mid z)] - \mathrm{KL}\left(q_\psi(z \mid x;\theta) \mid\mid p(z)\right)$$

⇒ SGD does not allow gradient to flow

Reparametrization : $$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

$$\nabla_\phi \mathrm{ELBO}(\phi,\theta) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[\nabla_\phi \log p_\theta(x|z) + \nabla_\phi \log q_\phi(z|x) - \nabla_\phi \log p(z)]$$

⇒ For rejection sampling, the number of random variable in the mapping is itself random

Reparametrization Trick

In Naesseth et al. (2017), derivation of a reparametrization trick for rejection sampling schemes based on the accepted sample density :

$$\pi(\varepsilon|\theta) \;=\; s(\varepsilon)\frac{g(h(\varepsilon,\theta)|\theta)}{r(h(\varepsilon,\theta)|\theta)}$$

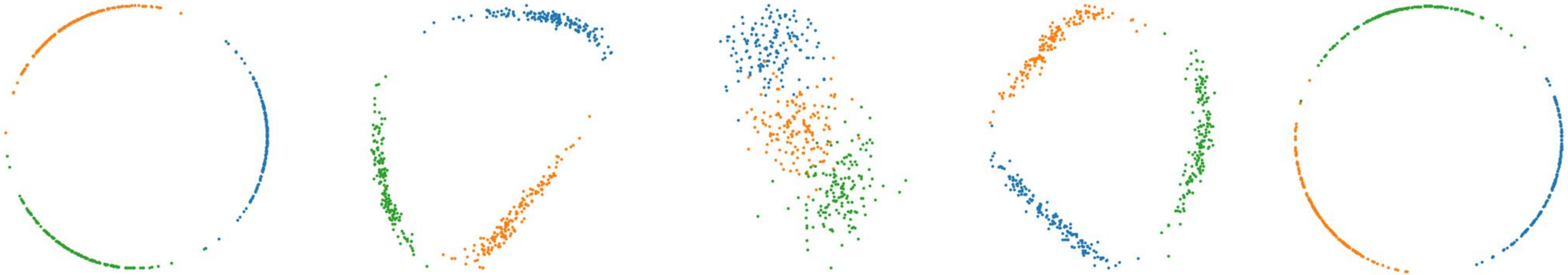
With log-derivative trick, leads to a 2-term gradient :

$$\nabla_\theta \mathbb{E}_{g(\omega|\theta)}[f(\omega)] = \mathbb{E}_{\pi(\varepsilon|\theta)}[\nabla_\theta f(h(\varepsilon,\theta))] + \mathbb{E}_{\pi(\varepsilon|\theta)}\left[f(h(\varepsilon,\theta))\nabla_\theta \log\frac{g(h(\varepsilon,\theta)|\theta)}{r(h(\varepsilon,\theta)|\theta)}\right]$$

In this paper, derived in the framework where there are 2 random variables

$\Rightarrow$ Same 2-term structure for the gradient, no dependance on mu

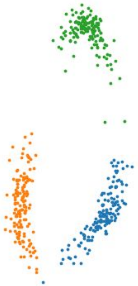
Experiments

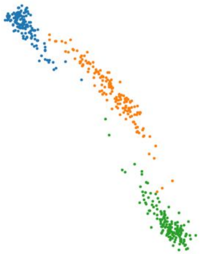


Original Auto-Encoder N-VAE, **β=10** N-VAE, β=0.1 S-VAE

Experiments

Original | Auto-Encoder | N-VAE, **β=10** | N-VAE, β=0.1 | S-VAE

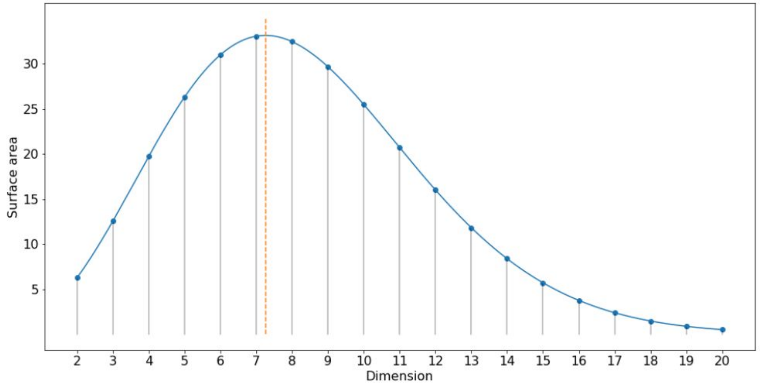




Final result highly dependent from the network initialization

Limitations

- Dimensionality
  ↳Vanishing surface
  ↳Similar performance in high dimension (~20)



- High variance term in reparametrization trick for rejection sampling

Image from the paper