

# 1 Tutorial on VAEs

Informellement : on cherche à apprendre une distribution. On se donne une loi de probabilité  $P(z)$  sur un espace latent  $\mathcal{Z}$ , et on veut optimiser les paramètres  $\theta$  d'une fonction  $f(z; \theta)$  de sorte à ce qu'en tirant  $z \sim P(z)$  et en calculant  $f(z; \theta)$ , on génère avec forte probabilité un échantillon qui ressemble aux éléments de notre dataset.

Cadre mathématique : maximisation de

$$P(X) = \int P(X|z; \theta) P(z) dz$$

avec  $P(X|z; \theta) = \mathcal{N}(X|f(z; \theta), \sigma^2 I)$ .

Idée clé des VAE : dans l'intégrale ci-dessus,  $P(X|z)$  n'a de contribution significative que pour un ensemble restreint de  $z \rightarrow$  intégrer selon  $Q(z|X)$  au lieu de  $P(z)$ .

Équation clé du VAE :

$$\log P(X) - \text{KL}[Q(z|X)||P(z|X)] = \mathbb{E}_{z \sim Q}[\log P(X|z)] - \text{KL}[Q(z|X)||P(z)] \quad (1)$$

À gauche, ce qu'on veut maximiser :  $\log P(X)$  d'une part, une divergence KL qui, si elle est minimisée, rend  $Q(z|X)$  proche de  $P(z|X)$ . À droite, quelque chose qu'on peut optimiser par SGD.

Choix usuel pour  $Q(z|X)$  :  $Q(z|X) = \mathcal{N}(z|\mu(X; \theta), \Sigma(X; \theta))$  où  $\mu$  et  $\Sigma$  sont implémentés comme des réseaux de neurones et  $\Sigma$  est contrainte à être diagonale.

La divergence KL  $\text{KL}[Q(z|X)||P(z)]$  a alors une forme close (cf le tutorial pour la formule, équation 7).

Pour le terme de gauche,  $\mathbb{E}_{z \sim Q}(\log P(X|z))$  : dans le cadre d'une SGD, échantillonner  $z$  une fois (selon  $Q(z|X)$ ) et utiliser  $P(X|z)$  pour approcher l'espérance.

Équation à optimiser en  $P$  et  $Q$  :

$$\mathbb{E}_{X \sim D}[\log P(X) - \text{KL}[Q(z|X)||P(z|X)]] = \mathbb{E}_{X \sim D}[\mathbb{E}_{z \sim Q}[\log P(X|z)] - \text{KL}[Q(z|X)||P(z)]] \quad (2)$$

où  $D$  est le dataset.

Problème : on ne peut pas faire rentrer le gradient directement dans  $\mathbb{E}_{z \sim Q}$ , car la distribution dépend de  $Q$  et donc des paramètres à optimiser. Solution : *reparameterization trick* : au lieu d'échantillonner selon  $Q(z|X) = \mathcal{N}(z|\mu(X; \theta), \Sigma(X; \theta))$  directement, échantillonner  $\epsilon \sim \mathcal{N}(0, I)$  puis calculer  $z = \mu(X; \theta) + \Sigma(X; \theta)^{1/2} \epsilon$ .

## 2 Hyperspherical Variational Auto-Encoders

### 2.1 Introduction

Exemple jouet qui montre selon les auteurs l'intérêt d'un  $\mathcal{S}$ -VAE : un cercle projeté dans  $\mathbb{R}^n$  par une certaine fonction  $f$  ; un autoencodeur découvre le cercle latent, pas un VAE.

Raison avancée : "a Gaussian prior is concentrated around the origin, while the KL-divergence tries to reconcile the differences between  $\mathcal{S}^1$  and  $\mathbb{R}^2$ " : ??

### 2.2 Variational Auto-Encoders

Problèmes de la distribution gaussienne :

- Basse dimension : concentration de la masse autour de l'origine, mauvais pour des distributions multimodales / avec plusieurs clusters.

- Haute dimension : distribution concentrée sur une hypersphère.

"The  $L_2$  norm [...] suffers from the curse of dimensionality" : ??

Discussion sur le mapping de variétés vers  $\mathcal{Z} \subset \mathbb{R}^D$  : soit  $\mathcal{M} \subset \mathbb{R}^M$  une variété ; en considérant un encodeur  $enc : \mathcal{M} \rightarrow \mathcal{Z}$ , sa corestriction à son image ne peut être un homéomorphisme que si  $D > M$  (sauf exceptions). Un VAE essaie de mapper la variété  $\mathcal{M}$  vers une distribution qui occupe tout l'espace latent  $\mathcal{Z}$ . Étant donné un encodeur  $enc$  qui induit un homéomorphisme entre  $\mathcal{M}$  et  $enc(\mathcal{M})$ , un VAE peut faire l'une de deux choses :

- soit le VAE ne fait que lisser  $enc(\mathcal{M})$  (pour occuper un peu tout l'espace?) et laisse une grande partie de  $\mathcal{Z}$  essentiellement vide, ce qui donne des mauvais samples,

- soit, lorsqu'on augmente la contribution de la divergence KL dans la quantité à maximiser (équation 1), il force l'encodeur à occuper tout l'espace latent, ce qui créerait de l'instabilité et des discontinuités.

Exemple de  $\mathcal{S}^1$ .

En général, il est difficile d'inférer la structure de la variété  $\mathcal{M}$  dans laquelle vit le dataset, mais les auteurs estiment qu'il est intéressant d'essayer de mapper les datasets vers des espaces autres que les  $\mathbb{R}^D$ .

## 2.3 Replacing Gaussian with von Mises-Fisher

On remplace le posterior par une von Mises-Fisher. Distribution sur une hypersphère  $\mathcal{S}^{m-1}$ . Paramètres  $\mu \in \mathcal{S}^{m-1}$ ,  $\kappa \geq 0$ .

Densité :

$$q(\mathbf{z}|\mu, \kappa) = \mathcal{C}_m(\kappa) \exp(\kappa \mu^\top \mathbf{z}) \quad (3)$$

où  $\mathcal{C}_m(\kappa)$  est une constante de normalisation.

On remplace le prior par la distribution uniforme sur  $\mathcal{S}^{m-1}$ . La divergence KL (équation 1, droite) a une forme close.

Échantillonnage d'une vMF : cf papier. On ne souffre pas du curse of dimensionality, car on fait un rejection sampling en 1D.