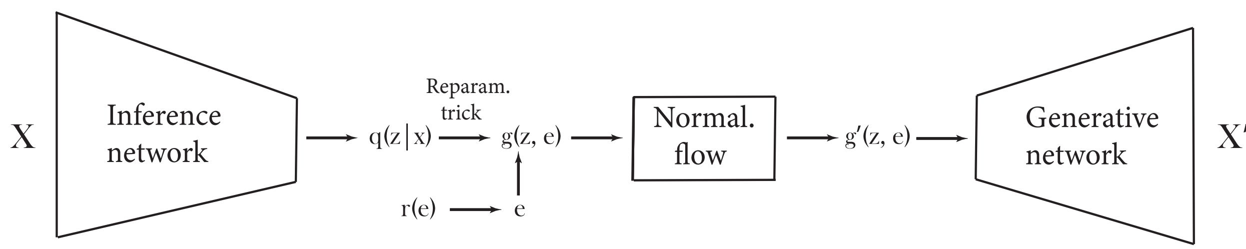
Basics of Variational Auto Encoder with Normalizing Flows

May 10, 2022

Useful papers - [RM15]

1 Introduction

The architecture of the variational auto encoder with normalizing flows is presented in the figure 1. The main goal of this model is to make latent posterior distribution more



expressive. ELBO looks like:

$$\mathcal{L}(\Theta, \Phi) = \mathbb{E}_{q(z|x,\Phi)} \log p(x|z, \Theta) - \mathbf{KL}(q(z|x, \Phi)||p(z)) \tag{1}$$

where $q(z|x, \Phi)$ is a latent posterior distribution. For analytically calculation of KL-divergence is it necessary to take simple posterior distribution, for instance gaussian $q(z|x, \Phi) = \mathcal{N}(z|\mu(x, \Phi), \sigma^2(x, \Phi) * \mathbf{I})$. However, might be that true posterior distribution $p(x|z)$ have more complex and non straightforward form. Thus, it is important to make approximation more expressive. Normalizing flows can be considered as efficient tool for that purpose. We consider $q(z|x, \Phi)$ as base distribution and pass flow throught it:

$$\log q^*(z^*|x, \Phi, \Lambda) = \log q(z|x, \Phi) - \log \left|\det \frac{\partial f(z, \Lambda)}{\partial z}\right| \tag{2}$$

where q^*, z^* are new distribution and variable, respectively; $z^* = f(z, \Lambda)$. $f(z, \Lambda)$ - flow model and could be presented as superposition of invertable transformations:

$$\begin{gathered} z^* = z_K = f_K \circ \ldots \circ f_2 \circ f_1(z) = \mathcal{F}(z) \\ \log q^*(z^*|x, \Phi, \Lambda) = \log q(z|x, \Phi) - \sum_{k=1}^{K} \log \left|\det \frac{\partial f_k}{\partial z_{k-1}}(z, \Lambda)\right| \end{gathered} \tag{3}$$

Let $q^*(z^*|x,\Phi,\Lambda)$ be new variational distribution, then ELBO:

$$\begin{gathered}
\mathcal{L}(\Theta,\Phi)=\mathbb{E}_{q^*(z^*|x,\Phi,\Lambda)}\log p(x|z^*,\Theta)-\mathbf{KL}(q^*(z^*|x,\Phi,\Lambda)||p(z))= \\
=\mathbb{E}_{q^*(z^*|x,\Phi,\Lambda)}\left[\log p(x,z^*|\Theta)-\log q^*(z^*|x,\Phi,\Lambda)\right]= \\
\left[\text{LOTUS}:\mathbb{E}_{p_y(y)}\,y=\mathbb{E}_{p_x(x)}\,g(x),\text{if } y=g(x)\right] \\
=\mathbb{E}_{q(z|x,\Phi)}\left[\log p(x,\mathcal{F}(z)|\Theta)-\log q^*(\mathcal{F}(z)|x,\Phi,\Lambda)\right]= \\
=\mathbb{E}_{q(z|x,\Phi)}\left[\log p(x,\mathcal{F}(z)|\Theta)-\log q(z|x,\Phi)+\sum_{k=1}^{K}\log\left|\det\frac{\partial f_k}{\partial z_{k-1}}(z_{k-1},\Lambda)\right|\right]
\end{gathered} \tag{4}$$

In VAE we have following assumptions

$$\begin{gathered}
p(x|z,\Theta)=\mathcal{N}(x|m(z,\Theta),c^2\mathbf{I}) \\
p(z)=\mathcal{N}(z|0,\mathbf{I}) \\
q(z|x,\Phi)=\mathcal{N}(z|\mu(x,\Phi),\sigma^2(x,\Phi)\mathbf{I})
\end{gathered} \tag{5}$$

Then

$$p(x,\mathcal{F}(z)|\Theta)=\mathcal{N}(x|m(\mathcal{F}(z),\Theta),c^2\mathbf{I})\mathcal{N}(\mathcal{F}(z)|0,\mathbf{I}) \tag{6}$$

We will consider residual normalizing flows

$$\begin{gathered}
f(z)=z+uh(w^Tz+b) \\
\Lambda=\{w\in\mathbb{R}^D,u\in\mathbb{R}^D,b\in\mathbb{R}\} \\
\left|\det\frac{\partial f}{\partial z}\right|=\left|1+u^Th'(w^Tz+b)w\right|
\end{gathered} \tag{7}$$

Then ELBO over whole dataset has following form

$$\begin{gathered}
\mathcal{L}(\Theta,\Phi)= \\
=\mathbb{E}_{p(x)}\,\mathbb{E}_{q(z|x,\Phi)}\left[\log p(x|\mathcal{F}_\Lambda(z),\Theta)+\log p(\mathcal{F}_\Lambda(z))-\log q(z|x,\Phi)+\sum_{k=1}^{K}\log\left|1+u_k^Th'(w_k^Tz_{k-1}+b_k)w_k\right|\right] \\
\log p(x|\mathcal{F}_\Lambda(z),\Theta)=const-||x-m(\mathcal{F}_\Lambda(z),\Theta)||^2 \\
\log p(\mathcal{F}_\Lambda(z))=const-||F_\Lambda(z)||^2 \\
\log q(z|x,\Phi)=\log\frac{1}{(2\pi)^{\frac{n}{2}}\prod_{i=1}^{n}\sigma_i}-\frac{1}{2}\left|\left|\frac{z-\mu(x,\Phi)}{\sigma(x,\Phi)}\right|\right|^2 \\
||\cdot||^2=\sum_{j=1}^{n}\cdot^2
\end{gathered} \tag{8}$$

For computing gradients reparameterization trick and M-C estimation is used

$$\begin{aligned} \Delta\Theta &= \nabla_{\Theta}\mathcal{L}(\Theta, \Phi) \\ \Delta(\Phi, \Lambda) &= \nabla_{\Phi,\Lambda}\mathcal{L}(\Theta, \Phi) \end{aligned} \tag{9}$$

References

[RM15] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.