(c) Powerful enough refers to if the neural network, parameterised by ϕ, underlying q_ϕ can accurately represent the true posterior distribution. i.e. if $q_\phi(Z|x) \approx p_\theta(Z|x)$ sufficiently well

$\exists \phi^*$ s.t. ... $q_\phi(z_k)$ ... $p_\theta(z_k)$ for $k=1,\dots,K$

In this case,

$$\underset{\theta,\phi}{\text{maximise}} \sum_{i=1}^N VLB^{(K)}_{\theta,\phi}(X_i)$$

closer to equality the more powerful q_ϕ is

$$\approx \underset{\theta}{\text{maximise}} \sum_{i=1}^N \mathbb{E}_{z_1,\dots,z_k \sim p_\theta(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^K \frac{p_\theta(x|z_k)p_z(z_k)}{p_\theta(z_k|x)}\right]$$

$$= \underset{\theta}{\text{maximise}} \sum_{i=1}^N \log\left(\frac{1}{K}\sum_{k=1}^K p_\theta(x)\right) = \underset{\theta}{\text{maximise}} \sum_{i=1}^N \log p_\theta(x) \quad \blacksquare$$

2.(a)

$$\log p_\theta(X_i) = \log\left(\mathbb{E}_{Z\sim r_\lambda(z)}\left[p_\theta(X_i|Z)\right]\right)$$

$$= \log\left(\mathbb{E}_{Z\sim q_\phi(z|X_i)}\left[\frac{p_\theta(X_i|Z)r_\lambda(Z)}{q_\phi(Z|X_i)}\right]\right)$$

$$\geq \mathbb{E}_{Z\sim q_\phi(z|X_i)}\left[\log\left(\frac{p_\theta(X_i|Z)r_\lambda(Z)}{q_\phi(Z|X_i)}\right)\right] \quad \text{by Jensen's inequality}$$

$$= VLB_{\theta,\phi,\lambda}(X_i) \quad \blacksquare$$

(b)

$$\underline{\nabla}\, VLB(X_i) = \left(\underline{\nabla}_\theta VLB(X_i), \underline{\nabla}_\phi VLB(X_i), \underline{\nabla}_\lambda VLB(X_i)\right)$$

$$\underline{\nabla}_\theta VLB(X_i) = \underline{\nabla}_\theta \int \log\left(\frac{p_\theta(X_i|z)r_\lambda(z)}{q_\phi(z|X_i)}\right) q_\phi(z|X_i)\,dz$$

$$= \int \underline{\nabla}_\theta\left(p_\theta(X_i|z)\right)\frac{1}{p_\theta(X_i|z)}\, q_\phi(z|X_i)\,dz + \underline{0}$$

$$= \mathbb{E}_{Z\sim q_\phi(z|X_i)}\left[\underline{\nabla}_\theta\left(\log(p_\theta(X_i|Z))\right)\right]$$

$$\underline{\nabla}_\phi VLB(X_i) = \mathbb{E}_{Z\sim q_\phi(z|X_i)}\left[\left(\underline{\nabla}_\phi \log q_\phi(\overset{Z|X_i}{Z})\right)\log\left(\frac{p_\theta(X_i|Z)r_\lambda(Z)}{q_\phi(Z|X_i)}\right)\right] \quad \text{by log-derivative trick for VAEs (HW10.1)}$$

$$\underline{\nabla}_\lambda VLB(X_i) = \mathbb{E}_{Z\sim q_\phi(z|X_i)}\left[\underline{\nabla}_\lambda\left(\log(r_\lambda(Z))\right)\right] \quad \text{by same logic as } \underline{\nabla}_\theta$$

So stochastic gradient of $VLB_{\theta,\phi,\lambda}(X_i)$ can be computed by:

$$\underline{\nabla}_{\theta,\phi,\lambda} VLB_{\theta,\phi,\lambda}(X_i) \approx \frac{1}{K}\sum_{k=1}^K\Big(\underline{\nabla}_\theta\left(\log(p_\theta(X_i|Z_{i,k}))\right), \left(\underline{\nabla}_\phi \log(\overset{Z_{i,k}|X_i}{Z_k})\right)\log\left(\frac{p_\theta(X_i|Z_k)r_\lambda(Z_k)}{q_\phi(Z_k|X_i)}\right),$$

$$\underline{\nabla}_\lambda\left(\log(r_\lambda(Z_{i,k}))\right)\Big) \quad \text{where } Z_k \sim q_\phi(z|X_i)$$

number of elements in a batch of SG calculation

$Z_{i,k}$ (sample for each X_i anew)

2.(c) ∇_θ and ∇_λ same as above (Expectation distribution q_ϕ doesn't depend on θ or λ so no need for the trick)

∇_ϕ evaluation changes if we use reparametrization vs. log-derivative before:

$$\nabla_\phi \, VLB(X_i) = \nabla_\phi \, \mathbb{E}_{Z \sim q_\phi(Z|X_i)}\left[\log\left(\frac{p_\theta(X_i|Z)\, r_\lambda(Z)}{q_\phi(Z|X_i)}\right)\right]$$

Redo 2.(c)

$$VLB_{\theta,\phi,\lambda}(X_i) = \underbrace{\mathbb{E}_{Z\sim q_\phi(Z|X_i)}\left[\log p_\theta(X_i|Z)\right]}_{①} \; \underbrace{\overbrace{- D_{KL}\left(q_\phi(Z|X_i) \| r_\lambda(Z)\right)}^{+\mathbb{E}\left[\log\left(\frac{r_\lambda}{q_\phi}\right)\right]}}_{②}$$

(slide 89)

- $① = \mathbb{E}_{Z\sim q_\phi(Z|X_i)}\left[\log\left((2\pi\sigma^2)^{-d/2} \exp\left(-\frac{1}{2}(X_i - f_\theta(Z))^T\left(\frac{1}{\sigma^2}I\right)(X_i - f_\theta(Z))\right)\right)\right]$ $\quad d = \dim(f_\theta)$

$$= -\frac{1}{2\sigma^2}\mathbb{E}_{Z\sim \mathcal{N}(\mu_\phi(X_i), \Sigma_\phi(X_i))}\left[\|X_i - f_\theta(Z)\|^2\right] - \frac{d}{2}\log(2\pi\sigma^2)$$

By reparametrization trick:

$$= -\frac{1}{2\sigma^2}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\left[\|X_i - f_\theta(\mu_\phi(X_i) + \Sigma_\phi^{1/2}(X_i)\varepsilon)\|^2\right] - \frac{d}{2}\log(2\pi\sigma^2)$$

$\downarrow$

$\rightarrow$ Apply ∇_ϕ : $-\frac{1}{2\sigma^2}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\left[\nabla_\phi \| \ldots \|^2\right]$ which can be evaluated + back-propagated

- $② = -\frac{1}{2}\left(tr\left([\mathrm{diag}(\lambda_2)]^{-1}\Sigma_\phi(X_i)\right) + (\lambda_1 - \mu_\phi(X_i))^T[\mathrm{diag}(\lambda_2)]^{-1}(\lambda_1 - \mu_\phi(X_i)) - k + \log\left(\frac{\det(\mathrm{diag}(\lambda_2))}{\det(\Sigma_\phi(X_i))}\right)\right)$

by HW9.5

$\rightarrow$ Gradient ∇_ϕ can then be calculated + back propagated