Optimization of the loss function

As discussed, $\theta^*, \phi^* = \underset{\theta, \phi}{\operatorname{argmin}} \; L(\theta, \phi)$

In variational Baysian method, this loss function is known as the variational lower bound or evidence lower bound (ELBO). This 'lower bound' part comes from the fact that KL divergence is always non-negative & thus $L(\theta, \phi)$ is the lower bound of $\log P_\theta(x)$.

Recall. (ELBO proof).

$$\log P_\theta(x) - D_{KL}\left[Q_\phi(z|x) \,\|\, P_\theta(z|x)\right] = -L(\theta, \phi)$$

Kumar

And we know $D_{KL}\left[Q_\phi(z|x) \,||\, P_\theta(z|x)\right] \geq 0$

As a result

$$L(\theta, \phi) \leq \log P_\theta(x)$$

Therefore minimizing loss, we are maximizing the lower bound of the probability of generating real data samples.

REPARAMETRIZATION TRICK

Recall: Needed during (Back Propagation)

$$L(\theta,\phi) = -E_{z\sim Q_\phi(z|x)}\left[\log\left(P_\theta(x|z)\right)\right] + \frac{1}{2}\sum_{K}\left[\exp\left(\Sigma_\phi(x)\right) + \mu_\phi^2(x) - 1 - \Sigma_\phi(x)\right]$$

$$\theta^*, \phi^* = \underset{\theta, \theta}{\operatorname{argmin}} \; L(\theta, \phi)$$

Alternate optimization principle

$$\theta^{**} = \{\nabla_\theta \{ L(\theta, \phi)\}\} \qquad \phi = \text{constant}$$

$$\phi^{**} = \nabla_\phi \{ L(\theta^{**}, \phi)\}$$

Optimization is carried out with respect to both θ & ϕ to learn $Q_\phi(z|x)$ & $P_\theta(x|z)$ at the same time i.e

$$\min_{\theta,\phi} L(\theta,\phi) = \min_{\theta,\phi}\left\{ -E_{z\sim Q_\phi(z|x)}\left[\log\left(P_\theta(x|z)\right)\right] + \frac{1}{2}\sum_{k}\left[\exp\left(\Sigma_\phi(x)\right) + \mu_\phi^2(x) - 1 - \Sigma_\phi(x)\right]\right\}$$

ALGORITHM

Alternate $i = 1, 2, \ldots n$

(A) $\hat{\theta}_i = \nabla_\theta L(\theta,\phi)$

$$\hat{\theta}_i = \nabla_\theta\left\{ -E_{z\sim Q_\phi(z|x)}\left[\log P_\theta(x|z)\right] + \frac{1}{2}\sum_{k}\left[\exp\left(\Sigma_\phi(x)\right) + \mu_\phi^2(x) - 1 - \Sigma_\phi(x)\right]\right\}$$

$$\hat{\theta}_i = \nabla_\theta L(\theta, \phi)$$

$$\cong \frac{1}{L} \sum_{l=1}^{L} \underline{\nabla_\theta \log P_\theta(x|z^{(l)})} \quad \left\{ \text{Monte Carlo estimate} \right\}.$$

where $z^{(l)} \sim Q_\phi(z|x)$

$$\hat{\theta}_i = \nabla_\theta L(\theta, \phi)$$

$$\hat{\theta}_i \cong \frac{1}{L} \sum_{l=1}^{L} \nabla_\theta \log P_\theta(x|z^{(l)}) \quad \left\{ \begin{array}{c} \text{Monte Carlo} \\ \text{estimate} \end{array} \right\}.$$

where $z^{(l)} \sim Q_\phi(z|x)$

$$\hat{\phi}_i = \nabla_\phi \left\{ L(\theta, \phi) \right\}$$

(B) $\hat{\phi}_i = \nabla_\phi \{ L(\theta, \phi) \}$

$$= \nabla_\phi \left\{ -E_{z \sim Q_\phi(z|x)} \left[\log P_\theta(x|z) \right] + \frac{1}{2} \sum_k \left\{ \exp(\Sigma_\phi(x)) + \mu_\phi^2(x) - 1 - \Sigma_\phi(x) \right\} \right\}$$

problem

This derivative ∇_ϕ is harder to estimate because ϕ appears in the distribution with respect to which expectation is taken. i.e $\nabla_\phi E_{Q_\phi(z/x)}[f(z)] \neq E_{Q_\phi(z/x)}[\nabla_\phi f(z)]$

If we can somehow rewrite this expectation in such a way the ϕ appears inside the expectation then we can push the gradient inside the expectation i.e if we can write

$$E_{Q_\phi(z/x)}[f(z)] = E_{p(\epsilon)}[f(g_\phi(\epsilon, x))]$$

Any linear transformation

such that $z = \boxed{g_\phi(\epsilon, x)}$ with $\epsilon \sim N(0,1)$

In our case $g_\phi(\epsilon, x) = \mu_\phi(x) + \epsilon \odot \Sigma_\phi^{1/2}(x) = z \sim N(\mu(x), \Sigma(x))$

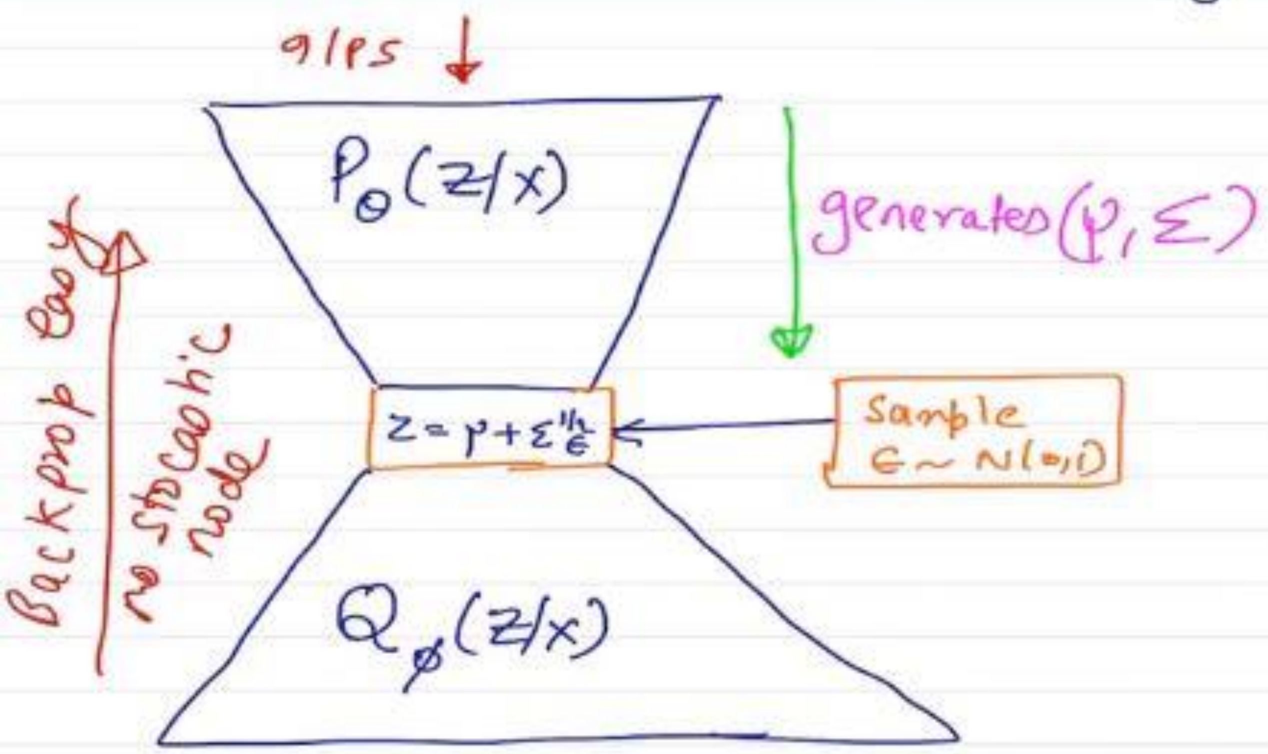
Here, $N(\mu(x), \Sigma(x))$ is obtained from $N(0,1)$ using above linear transformation.

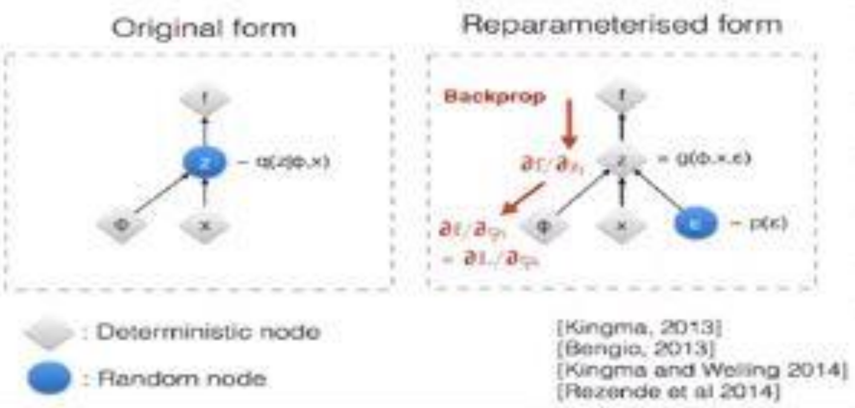
Instead of sampling $z \sim Q_\phi(z/x)$, we sample from $N(0,1)$

i.e $\epsilon \sim N(0,1) \Rightarrow \epsilon \sim p(\epsilon)$

& then linear transform using $z = \mu_\phi(x) + \epsilon \odot \Sigma_\phi^{1/2}(x)$

to realise $N(\mu(x), \Sigma(x))$ defined earlier





Kumar

$$\hat{\phi}_i = \nabla_\phi \{ L(\theta, \phi) \}$$

$$= \nabla_\phi \left\{ - E_{z \sim Q_\phi(z|x)} \left[\log P_\theta (x|z) \right] + \frac{1}{2} \sum_k \left[\exp(\Sigma_\phi(x)) + \mu_\phi^2(x) - 1 - \Sigma_\phi(x) \right] \right\}$$

problem

Repeated

modified

$$\hat{\phi}_i = \nabla_\phi \{ L(\theta, \phi) \}$$

$$= \nabla_\phi \left\{ - E_{z^{(l)} \sim p(\epsilon)} \left[\log P_\theta (x|z^{(l)}) \right] + \frac{1}{2} \sum_k \left[\exp(\Sigma_\phi(x)) + \mu_\phi^2(x) - 1 - \Sigma_\phi(x) \right] \right\}$$

$$\hat{\phi}_i = -E_{z^l \sim p(\epsilon)}\left[\nabla_\phi\left(\log p_\theta(x|z^l)\right)\right] +$$

$$\nabla_\phi\left[\frac{1}{2}\left(\sum_K \left[\exp(\Sigma_\phi(x)) + \mu_\phi^2(x) - 1 - \Sigma_\phi(x)\right]\right)\right]$$

$$= -\frac{1}{S}\sum_{s=1}^{S} \cdot \left[\log p_\theta(x|z^{(l)})\right] +$$

Monte-Carlo estimate of Expectation.

where $z^{(l)} = \mu_\phi(x) + \epsilon \odot \sigma_\phi(x)$ & $\epsilon^{(l)} \sim N(0,1)$