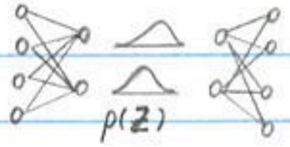
p1. <u>Evidence Lower Bound (ELBO) for VAE</u>

<u>Goal:</u> A VAE is an autoencoder so we need our network to reconstruct the data. However, we also need our network to learn to map the latent space to a choosen prior. Thus, we need an objective function that does both of these things.

Let:



$q_\theta(z|x)$ $\quad$ $p_\phi(x|z)$

Recall: Baye's Rule, $\underbrace{p(z|x)}_{\text{posterior}} = \underbrace{p(z)}_{\text{prior}} \cdot \dfrac{\overbrace{p(x|z)}^{\text{likelihood}}}{\underbrace{\int p(x|z)\,p(z)\,dz}_{\text{margin likelihood}}}$

<u>ELBO for VAE:</u> 1. Let $p(z|x;)$ be a choosen prior distribution to express latent space.

2. To make the encoder learn to map to our chosen prior, we can minimize, $D_{KL}(q_\theta(z|x;)\,||\,p(z|x;))$

3. We shall proceed to show that $D_{KL}(q_\theta(z|x;)\,||\,p(z|x;))$ can be used to construct an objective function that maximizes the log-likelihood of our data (the goal of any probabilistic model) by optimizing a lower-bound that takes into account the VAE's objectives of mapping to a choosen prior for the latent and reconstruction of data.

p2.

4. $$D_{KL}(q_\theta(z|X_i)\,||\,p(z|X_i)) = -\int q_\theta(z|X_i)\log\left(\frac{p(z|X_i)}{q_\theta(z|X_i)}\right)dz \geq 0$$

$$= -\int q_\theta(z|X_i)\log\left(\frac{p_\phi(X_i|z)\,p(z)}{q_\theta(z|X_i)\,p(X_i)}\right)dz \geq 0$$ // apply bayes to $p(z|X_i)$

$$= -\int q_\theta(z|X_i)\left[\log\left(\frac{p_\phi(X_i|z)\,p(z)}{q_\theta(z|X_i)}\right) - \log(p(X_i))\right]dz \geq 0$$ // $\log(\frac{a}{b}) = \log(a) - \log(b)$

$$= \left(-\int q_\theta(z|X_i)\log\left(\frac{p_\phi(X_i|z)\,p(z)}{q_\theta(z|X_i)}\right)dz\right) + \left(\log(p(X_i))\underbrace{\int q_\theta(z|X_i)\,dz}_{\text{integrates to } 1}\right) \geq 0$$

$$= \log(p(X_i)) - \int q_\theta(z|X_i)\log\left(\frac{p_\phi(X_i|z)\,p(z)}{q_\theta(z|X_i)}\right)dz \geq 0$$

$$\log(p(X_i)) \geq \int q_\theta(z|X_i)\log\left(\frac{p_\phi(X_i|z)\,p(z)}{q_\theta(z|X_i)}\right)dz$$

$$\geq \int q_\theta(z|X_i)\log\left(\frac{p(z)}{q_\theta(z|X_i)}\right)dz + \int q_\theta(z|X_i)\log(p_\phi(X_i|z))\,dz$$

$$\geq \underbrace{-D_{KL}(q_\theta(z|X_i)\,||\,p(z)) + \mathbb{E}_{\sim q_\theta(z|X_i)}\left[\log(p_\phi(X_i|z))\right]}_{\text{"ELBO"}}$$

5. $$\log(p(X)) \geq \underbrace{-D_{KL}(q_\theta(z|X_i)\,||\,p(z))} + \underbrace{\mathbb{E}_{\sim q_\theta(z|X_i)}\left[\log(p_\phi(X_i|z))\right]}$$

- This term causes the encoder to map to a chosen prior, $p(z)$.

- This term causes the decoder to map the latent back to our data's space.

- Also, by maximizing ELBO, we maximize the likelihood of our data being observed by our network. Thus, our network learns in respect to our data. $\square$

Using ELBO for a VAE in practice: we have shown that maximizing ELBO fits a VAE yet, in practice, we must choose a prior, $p(z)$, for the latent to map the data too. We then need to manipulate ELBO to learn the parameters of the chosen prior.

Closed Form VAE Loss, Gaussian Latent: we shall show ELBO for when the prior, $p(z)$, is gaussian.

1. Let $Z \sim N(\mu, \sigma^2)$, then $p(z) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(\frac{-(X-\mu_p)^2}{2\sigma_p^2}\right)$ and $q_\theta(z|X_i) = \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(\frac{-(X-\mu_q)}{2\sigma_q^2}\right)^2$

2. Plug in $p(z)$ and $q_\theta(z|X_i)$ into the KL term of ELBO,

$$-D_{KL}(q_\theta(z|X_i) \| p(z)) = \int q_\theta(z|X_i) \log\left(\frac{q_\theta(z|X_i)}{p(z)}\right) dz$$

$$= \int \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(\frac{-(X-\mu_q)^2}{2\sigma_q^2}\right) \log\left(\frac{\frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(\frac{-(X-\mu_p)^2}{2\sigma_p^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(\frac{-(X-\mu_q)}{2\sigma_q^2}\right)}\right) dz$$

3. We will work the current form of D_{KL} to be in terms of $E(\cdot)$ and $V(\cdot)$

$$-D_{KL}(q_\theta(z|X_i) \| p(z)) = \int \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(\frac{-(X-\mu_q)^2}{2\sigma_q^2}\right) \cdot \left[\left(-\tfrac{1}{2}\right) \log(2\pi) - \log(\sigma_p) - \frac{(X-\mu_p)^2}{2\sigma_p^2} + \left(\tfrac{1}{2}\right) \log(2\pi) + \log(\sigma_q) + \frac{(X-\mu_q)^2}{2\sigma_q^2} \right] dz$$

p4.

$$= \frac{1}{\sqrt{2\pi\sigma_q^2}} \int \exp\left(\frac{-(x-\mu_q)^2}{2\sigma_q^2}\right) \left[-\log(\sigma_p) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \log(\sigma_q) + (1/2)\log(2\pi) + \log(\sigma_q) + \frac{(x-\mu_q)^2}{2\sigma_q^2} \right] dz$$

$$= \frac{1}{\sqrt{2\pi\sigma_q^2}} \int \exp\left(\frac{-(x-\mu_q)^2}{2\sigma_q^2}\right) \left(\log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2} \right) dz$$

$$\Rightarrow \quad -D_{KL}(q_\theta(z|x_i) \| p(z)) = E_q\left[\log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2} \right]$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) + E_q\left[\frac{-(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2} \right]$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2} E_q[(x-\mu_p)^2] + \frac{1}{2\sigma_q^2} E_q[(x-\mu_q)^2]$$

4. Recall $V(\cdot) = \sigma^2 = E[(x-\mu)^2]$, thus

$$-D_{KL}(q_\theta(z|x_i) \| p(z)) = \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2} E_q[(x-\mu_p)^2] + \frac{\sigma_q^2}{2\sigma_q^2}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2} E_q[(x-\mu_p)^2] + \frac{1}{2}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2} E_q[(\underbrace{x-\mu_q}_{a} + \underbrace{\mu_q-\mu_p}_{b})^2] + \frac{1}{2}$$

recall, $(a+b)^2 = a^2 + 2ab + b^2$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2} E_q[(x-\mu_q)^2 + 2(x-\mu_q)(\mu_q-\mu_p) + (\mu_q-\mu_p)^2] + \frac{1}{2}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{\sigma_q^2 + (\mu_q-\mu_p)^2}{2\sigma_p^2} + \frac{1}{2}$$

p5.

5. Assume prior, $p(z)$, is $N(0,1)$. Thus, $\sigma_p = 1$ and $\mu_p = 0$.

$$\therefore \quad -D_{KL}(q_\theta(z|X_i) \| p(z)) = \log(\sigma_q) - \frac{\sigma_q^2 + \mu_q^2}{2} + \frac{1}{2}$$

$$= (1/2)[1 + \log(\sigma_p^2) - \sigma_q^2 - \mu_q^2]$$

6. Thus, ELBO with a standard normal prior is

$$\log(p(X_i)) \geq \tfrac{1}{2} \cdot [1 + \log(\sigma_j^2) - \sigma_j^2 - \mu_j^2] + \mathbb{E}_{\sim q_\theta(z|X_i)}[\log(p_\phi(X_i|z))]$$

For a training batch, ELBO with $N(0,1)$ prior is

$$G = \sum_{j=1}^{J} (1/2)[\log(\sigma_j^2) - \sigma_j^2 - \mu_j^2 + 1] + \frac{1}{L} \sum_{\ell} \mathbb{E}_{\sim q_\theta(z|X_i)}[\log(p(X_i|z^{(i,\ell)}))]$$

where J is the dimension of the latent vector z, and L is number of samples drawn according to reparametrization trick.

Re-parametrization Trick : In ELBO, $\mathbb{E}_{\sim q_\theta(z|X_i)}[\log(p_\phi(X_i|z))]$ is not differentiable. Thus, SGD does not work. By using Stochastic Gradient Variational Bayes (SGVB), we can approximate $\mathbb{E}_{\sim q_\theta(z|X_i)}[\log(p_\phi(X_i|z))]$.

To do so, we will reparametrize the random variable $\tilde{z} \sim q_\phi(z|X)$ using a differentiable function $g_\phi(\epsilon, X)$ where ϵ is a noise variable

$$\tilde{z} = g_\phi(\epsilon, X) \quad \text{with} \quad \epsilon \sim p(\epsilon)$$

We can form Monte Carlo estimates of expectations of some function $f(z)$ w.r.t. $q_\phi(z|x)$ as follows:

$$\mathbb{E}_{q_\phi(z|x_i)}[f(z)] = \mathbb{E}_{p(\varepsilon)}[f(g_\phi(\varepsilon, x_i))] \simeq \frac{1}{L}\sum_{l=1}^{L} f(g_\phi(\varepsilon^{(l)}, x^{(i)}))$$

where $\varepsilon^{(l)} \sim p(\varepsilon)$

Applying to lower bound $\mathbb{E}_{\sim q_\theta(z|x_i)}[\log(p(x_i|z^{(i,l)}))]$

$$\tilde{\mathcal{L}}^A(\theta, \phi; x^{(i)}) \simeq \mathcal{L}(\theta, \phi; x^{(i)})$$

$$\tilde{\mathcal{L}}^A(\theta, \phi; x^{(i)}) = \frac{1}{L}\sum_{l=1}^{L} \log(p_\theta(x^{(i)}, z^{(i,l)})) - \log(q_\phi(z^{(i,l)}| x^{(i)})$$

where $z^{(i,l)} = g_\phi(\varepsilon^{(i,l)}, x^{(i)})$ and $\varepsilon^{(l)} \sim p(\varepsilon)$

Proof: 1. Given the deterministic mapping $z = g_\phi(\varepsilon, x)$ we know that $q_\phi(z|x)\prod_i dz_i = p(\varepsilon)\prod_i d\varepsilon_i$

2. $\therefore \int q_\phi(z|x) f(z)\,dz = \int p(\varepsilon) f(z)\,d\varepsilon = \int p(\varepsilon) f(g_\phi(\varepsilon, x))\,d\varepsilon$

3. Thus, a differentiable estimator can be made $\int q_\phi(z|x) f(z)\,dz \simeq \frac{1}{L}\sum_{l=1}^{L} f(g_\phi(x, \varepsilon^{(l)}))$ where $\varepsilon^{(l)} \sim p(\varepsilon)$

VAE ELBO Re-param Gauss:

$$\log(p(x_i)) \geq (\tfrac{1}{2})[1 + \log(\sigma_j^2) - \sigma_j^2 - \mu_j^2] + \mathbb{E}_{\sim q_\theta(z|x_i)}[\log(p_\phi(x_i|z_i))]$$

where $z_j = \mu_j + \sigma_j \odot \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, 1)$