- Evidence Lower Bound

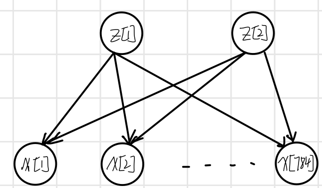
$$\begin{aligned}
\log P(X;\theta) &= \sum_i \log \sum_{z_i \in Val(z_i)} P(x_i, z_i;\theta) \\
&= \sum_i \log \sum_{z_i \in Val(z_i)} q_i(z_i) \frac{P(x_i, z_i;\theta)}{q_i(z_i)} \\
&\geq \sum_i \sum_{z_i \in Val(z_i)} q_i(z_i) \log \frac{P(x_i, z_i;\theta)}{q_i(z_i)} \\
&= \sum_i \left[H(q_i) + E_{z_i \sim q_i} \log P(x_i, z_i;\theta) \right] \\
&= \sum_i \mathcal{L}(x_i)
\end{aligned}$$

$$\begin{aligned}
\log P(X;\theta) &= \sum_i \log p(x_i;\theta) \\
&= \sum_i \sum_{z_i} q_i(z_i) \log p(x_i;\theta) \\
&= \sum_i \sum_{z_i} q_i(z_i) \left[\log \frac{P(x_i, z_i;\theta)}{P(z_i|x_i;\theta)} + \log q_i(z_i) - \log q_i(z_i) \right] \\
&= \sum_i \sum_{z_i} q_i(z_i) \left[\log \frac{q_i(z_i)}{P(z_i|x_i;\theta)} + \log P(x_i, z_i;\theta) - \log q_i(z_i) \right] \\
&= \sum_i \left[KL\left(q_i(z_i) \,\|\, P(z_i|x_i;\theta)\right) + \mathcal{L}(x_i) \right]
\end{aligned}$$

when $q_i(z_i) = P(z_i|x_i;\theta)$, the lower bound is tight

learn a neural net $q(z|x;\phi)$ to represent $q_i(z_i) \approx P(z_i|x_i;\theta)$

$$\begin{aligned}
\mathcal{L}(x_i) &= H(q_i) + E_{z_i \sim q_i}\left[\log P(x_i, z_i;\theta) \right] \\
&= H(q_i) + E_{z_i \sim q_i}\left[\log P(z_i) + \log P(x_i|z_i;\theta) + \log q(z_i|x_i;\phi) - \log q(z_i|x_i;\phi) \right] \\
&= E_{z_i \sim q_i}\left[\log P(z_i) + \log P(x_i|z_i;\theta) - \log q(z_i|x_i;\phi) \right] \\
&= -KL\left(q(z_i|x_i;\phi) \,\|\, P(z_i) \right) + E_{z_i \sim q_i}\left[\log P(x_i|z_i;\theta) \right]
\end{aligned}$$



$$\begin{aligned}
\log P(X;\theta) &\geq \sum_i \mathcal{L}(x_i) \\
&= \sum_i \Big[\underbrace{E_{z_i \sim q_i} \log P(x_i|z_i;\theta)}_{\text{approximate log likelihood}} - \underbrace{KL\left(q(z_i|x_i;\phi) \,\|\, P(z_i) \right)}_{\text{evaluate / approximate KL}} \Big]
\end{aligned}$$

· KL In Gaussian VAE

in Gaussian VAE, assume $P(z) \sim \mathcal{N}(0, I)$, $\quad q(z_i|x_i;\phi) \sim \mathcal{N}(\mu_\phi(x_i), \mathrm{diag}(\sigma_\phi^2(x_i)))$

$$P(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

$$\log P(x;\mu,\Sigma) = -\frac{1}{2}\log(2\pi)^n|\Sigma| - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

$$KL\left(\mathcal{N}(\mu_1,\Sigma_1)\|\mathcal{N}(\mu_2,\Sigma_2)\right) = \int_{\mathbb{R}^n} \mathcal{N}(x;\mu_1,\Sigma_1)\cdot\left[-\frac{1}{2}\log 2\pi^n|\Sigma_1| + \frac{1}{2}\log 2\pi^n|\Sigma_2| - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right] dx$$

$$= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}\int_{\mathbb{R}^n}\mathcal{N}(x;\mu_1,\Sigma_1)\left\langle (x-\mu_1)(x-\mu_1)^T, \Sigma_1^{-1}\right\rangle dx$$

$$+ \frac{1}{2}\int_{\mathbb{R}^n}\mathcal{N}(x;\mu_1,\Sigma_1)\ (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\ dx$$

$$= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}\left\langle E_{x_i}\left[(x-\mu_1)(x-\mu_1)^T\right], \Sigma_1^{-1}\right\rangle + \frac{1}{2}\int_{\mathbb{R}^n}\mathcal{N}(x;\mu_1,\Sigma_1)\ (x_1-\mu_1+\mu_1-\mu_2)\Sigma_2^{-1}(x_1-\mu_1+\mu_1-\mu_2)\ dx$$

$$= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{n}{2} + \frac{1}{2}\int_{\mathbb{R}^n}\mathcal{N}(x;\mu_1,\Sigma_1)\left[\left\langle (x-\mu_1)(x-\mu_1)^T, \Sigma_2^{-1}\right\rangle + 2(x-\mu_1)^T\Sigma_2^{-1}(x-\mu_2) + (\mu_1-\mu_2)\Sigma_2^{-1}(\mu_1-\mu_2)\right] dx$$

$$= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{n}{2} + \frac{1}{2}\left[\left\langle \Sigma_1, \Sigma_2^{-1}\right\rangle + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right]$$

when $\Sigma_1 = \mathrm{diag}(\sigma_1^2)$, $\Sigma_2 = \mathrm{diag}(\sigma_2^2)$

$$KL\left(\mathcal{N}(\mu_1,\Sigma_1)\|\mathcal{N}(\mu_2,\Sigma_2)\right) = \sum_i \frac{1}{2}\left[\log\frac{\sigma_{2i}^2}{\sigma_{1i}^2} - 1 + \frac{\sigma_{1i}^2}{\sigma_{2i}^2} + \frac{(\mu_{1i}-\mu_{2i})^2}{\sigma_{2i}^2}\right]$$

- KL in MGVAE

in MGVAE, assume: $P(z) = \frac{1}{K} \sum_K \mathcal{N}(z; \mu_k, \mathrm{diag}(\sigma_k^2))$

$q(z|x;\phi) = \mathcal{N}(z; \mu_\phi(x), \mathrm{diag}(\sigma_\phi^2(x)))$

$$KL\left(q(z_i|x_i;\phi) \,||\, P(z_i) \right) = E_{z_i \sim q(z_i|x_i;\phi)}\left[\log q(z_i|x_i;\phi) - \log P(z_i) \right]$$

$$\approx \frac{1}{K} \sum_K \left[\log q(z_i^{(k)}|x_i;\phi) - \log P(z_i^{(k)}) \right]$$

$$= \frac{1}{K} \sum_K \left[\log \mathcal{N}(z_i^{(k)}; \mu_\phi(x_i), \sigma_\phi^2(x_i)) - \log \sum_j w_j \mathcal{N}(z_i^{(k)}; \mu_j, \sigma_j^2) \right]$$

$$\log \mathcal{N}(x; \mu; \mathrm{diag}(\sigma)) = \log \left[\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right) \right]$$

$$= \sum_i \left[-\log \sqrt{2\pi} - \log \sigma_i - (x_i - \mu_i)^2 / 2\sigma_i^2 \right]$$

$$\left.\begin{aligned} & \log \sum_j w_j \exp \log \mathcal{N}_j \\ &= \log \left[\exp(\log \mathcal{N}_{max}) \cdot \sum_j w_j \exp(\log \mathcal{N}_j - \log \mathcal{N}_{max}) \right] \\ &= \log \mathcal{N}_{max} + \log \sum_j w_j \exp(\log \mathcal{N}_j - \log \mathcal{N}_{max}) \end{aligned}\right\}$$

numerical stable computation of log-weight-sum-exp

· Importance Weighting

$$\log P(X;\theta) = \sum_i \log \sum_{z_i} P(x_i, z_i; \theta)$$

$$= \sum_i \log \sum_{z_i} q_i(z_i|x_i;\phi) \cdot \frac{P(x_i, z_i;\theta)}{q_i(z_i|x_i;\phi)}$$

$$= \sum_i \log E_{z_i \sim q_i(z_i|x_i;\phi)} \frac{P(x_i, z_i;\theta)}{q_i(z_i|x_i;\phi)}$$

$$= \sum_i \log \left[\frac{1}{K} \sum_k E_{z_i^{(k)} \sim q_i} \frac{P(x_i, z_i^{(k)};\theta)}{q_i(z_i^{(k)}|x_i;\theta)} \right]$$

$$= \sum_i \log \left[E_{z_i^{(1)} \cdots z_i^{(k)} \sim q_i} \frac{1}{K} \sum_k \frac{P(x_i, z_i^{(k)};\theta)}{q_i(z_i^{(k)}|x_i;\theta)} \right]$$

$$\geq \sum_i E_{z_i^{(1)} \cdots z_i^{(k)} \sim q_i} \left[\log \frac{1}{K} \sum_k \frac{P(x_i, z_i^{(k)};\theta)}{q_i(z_i^{(k)}|x_i;\theta)} \right] = \sum_i E_{z_i^{(1)} \cdots z_i^{(k)} \sim q_i} \left[\log \frac{1}{K} \sum_k \frac{P(x_i|z_i^{(k)};\theta) \cdot P(z_i^{(k)})}{q_i(z_i^{(k)}|x_i;\theta)} \right]$$

importance weighting lower bound

$$\geq \sum_i E_{z_i^{(1)} \cdots z_i^{(k)} \sim q_i} \left[\frac{1}{K} \sum_k \log \frac{P(x_i, z_i^{(k)};\theta)}{q_i(z_i^{(k)}|x_i;\theta)} \right]$$

$$= \sum_i E_{z_i \sim q_i} \left[\log \frac{P(x_i, z_i;\theta)}{q_i(z_i|x_i;\theta)} \right]$$

$$= \sum_i E_{z_i \sim q_i} \left[\log \frac{P(z_i)}{q_i(z_i|x_i;\theta)} + \log P(x_i|z_i^{(k)}) \right]$$

$$= \sum_i \left[KL\left(q_i(z_i|x_i;\theta) \,||\, P(z_i) \right) + E_{z_i \sim q_i}\left[\log P(x_i|z_i^{(k)}) \right] \right] = \text{evidence lower bound}$$

$$\log P(X;\theta) \geq \text{importance weighting lower bound} \geq \text{evidence lower bound}$$

$$\log \frac{1}{K} \sum_k \exp \log \frac{P(x_i|z_i^{(k)};\theta) \cdot P(z_i^{(k)})}{q_i(z_i^{(k)}|x_i;\theta)}$$

$$= \log \frac{1}{K} \sum_k \exp \left[\log P(x_i|z_i^{(k)};\theta) + \log P(z_i^{(k)}) - \log q_i(z_i^{(k)}|x_i;\theta) \right]$$

$$= \log \frac{1}{K} \sum_k \exp LP(k)$$

$$= \log \left[\exp(LP_{max}) \cdot \frac{1}{K} \sum_k \exp\left(LP(k) - LP(max) \right) \right]$$

$$= LP[max] + \log \frac{1}{K} \sum_k \exp\left(LP(k) - LP(max) \right)$$

· Semi-Supervised VAE

Assume $P(z) = \mathcal{N}(z; 0, I)$

$p(y) = \frac{1}{10}$

$p(x|y,z) = \text{Bernoulli}(x \mid f_\theta(y,z))$ (for MNIST data set)

X_l = labeled data $\quad$ X_u = unlabled data

evidence lower bound for unlabeled data

$$l(\theta) = \sum_{x_i \in X_u} \log p(x_i;\theta) + \sum_{x_i \in X_l} P(x_i, y_i;\theta)$$

$$\sum_{x_i \in X_u} \log p(x_i;\theta)$$

$$= \sum_{x_i \in X_u} \log \sum_{z_i, y_i} p(x_i, y_i, z_i;\theta)$$

$$= \sum_{x_i \in X_u} \log \sum_{z_i, y_i} q_i(z_i, y_i) \frac{p(x_i, y_i, z_i;\theta)}{q_i(z_i, y_i)}$$

$$\geq \sum_{x_i \in X_u} \sum_{z_i, y_i} q_i(z_i, y_i) \log \frac{p(x_i, y_i, z_i;\theta)}{q_i(z_i, y_i)}$$

when $q_i(z_i, y_i) = P(z_i, y_i | x_i;\theta)$, the lower bound is tight

learn a neural net $\begin{cases} q(y|x;\phi) = \text{Categorial}(y \mid f_\phi(x)) \\ q(z|x,y) = \mathcal{N}(z \mid \mu_\phi(x,y), \text{diag}(\sigma^2(x,y))) \end{cases}$

to represent $q(z_i, y_i | x_i) = q(y_i|x_i;\phi) \cdot q(z_i|x_i, y_i)$

$$= \sum_{x_i \in X_u} \sum_{z_i, y_i} q(y_i|x_i;\phi) \cdot q(z_i|x_i,y_i;\phi) \cdot \log \frac{p(z_i)\, p(y_i) \cdot \overbrace{p(x_i|y_i,z_i;\theta)}^{\text{decoder}}}{\underbrace{q(y_i|x_i;\phi)}_{\text{classifier}}\, \underbrace{q(z_i|x_i,y_i;\phi)}_{\text{encoder}}}$$

$$= \sum_{x_i \in X_u} E_{y_i \sim q(y_i|x_i;\phi)} E_{z_i \sim q(z_i|x_i,y_i;\phi)} \left[\log \frac{p(z_i)}{q(z_i|x_i,y_i;\phi)} + \log p(x_i|y_i,z_i;\theta) + \log \frac{p(y_i)}{q(y_i|x_i;\phi)} \right]$$

$$= \sum_{x_i \in X_u} \left\{ E_{y_i \sim q(y_i|x_i;\phi)} \log \frac{p(y_i)}{q(y_i|x_i;\phi)} + E_{y_i \sim q(y_i|x_i;\phi)} E_{z_i \sim q(z_i|x_i,y_i;\phi)} \left[\log \frac{p(z_i)}{q(z_i|x_i,y_i;\phi)} + \log p(x_i|y_i,z_i;\theta) \right] \right\}$$

$$= \sum_{x_i \in X_u} \left[-KL(q(y_i|x_i;\phi) \,\|\, p(y_i)) - E_{y_i \sim q(y_i|x_i;\phi)} \left[KL(q(z_i|x_i,y_i;\phi) \,\|\, p(z_i)) \right] + E_{y_i \sim q(y_i|x_i;\phi)} E_{z_i \sim q(z_i|x_i,y_i;\phi)} \log p(x_i|y_i,z_i;\theta) \right]$$