$x^1, x^2, \cdots, x^m$

$$P_\theta(x) = \int P_\theta(x, z)\, dz$$

$$= \int P_\theta(x|z)\, p(z)\, dz$$

log-likelihood

$$\ell\ell(\theta) = \sum_{i=1}^{m} \log P_\theta(x^i) = \sum_{i=1}^{m} \log \int P_\theta(x^i, z)\, dz$$

→ Hard!

ELBO

$$\mathcal{L}(q_i, x^i, \theta) = \sum q_i(z|x^i) \log \frac{P_\theta(x^i, z)}{q_i(z|x^i)}$$

$$= \log P_\theta(x^i) - KL\left(q_i(z|x^i) \,\|\, P_\theta(z|x^i)\right)$$

$$\max_\theta \ell\ell(\theta) \Longrightarrow \max_{\theta, q_1, q_2, \cdots, q_m} \sum_{i=1}^{m} \mathcal{L}(q_i, x^i, \theta)$$

$$= \max_\theta \sum_{i=1}^{m} \log P_\theta(x^i)$$

start from some $\theta, q_1, q_2, \cdots q_m$

Loop

$$\nabla_\theta = \frac{\partial}{\partial \theta} \sum_{i=1}^{m} \mathcal{L}(q_i, x^i, \theta)$$

$$\theta \leftarrow \theta + \lambda \nabla_\theta$$

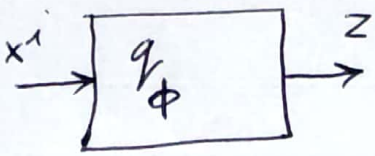
for $i = 1 \cdots m$:

$$q_i \leftarrow \operatorname*{argmax}_q \mathcal{L}(q, x^i, \theta)$$

what if we have millions of data?

solution 1 $q_i(z|x^i) \longrightarrow q(z)$ ✗

solution 2: $q_i(z|x^i) \Rightarrow q_\phi(z|x^i)$



Amortized Variational Inference

Amortized VI

start from some θ, ϕ

Loop

$$\nabla_\theta = \frac{\partial}{\partial \theta} \sum_{i=1}^{m} \mathcal{L}(x^i, \phi, \theta)$$

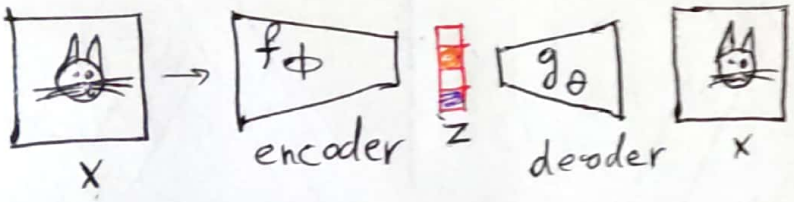
$$\nabla_\phi = \frac{\partial}{\partial \phi} \sum_{i=1}^{m} \mathcal{L}(x^i, \phi, \theta)$$

$$\theta \leftarrow \theta + \lambda \nabla_\theta$$

$$\phi \leftarrow \phi + \lambda \nabla_\phi$$

$$\mathcal{L}(x^i, \phi, \theta) = \sum_z q_\phi(z|x^i) \log \frac{p_\theta(z, x^i)}{q_\phi(z|x^i)}$$

Auto-encoders



$$z = f_\phi(x)$$

$$y = g_\theta(z)$$

find ϕ, θ such that $\|y - x\|$ is small

$$\min_{\theta, \phi} \sum_{i=1}^{m} \left\| g_\theta(f_\phi(x^i)) - x^i \right\|^2$$

1- data compression

2- dimentionality reduction

3- feature extraction (unsupervised), representation learning

4- semi-supervised learning

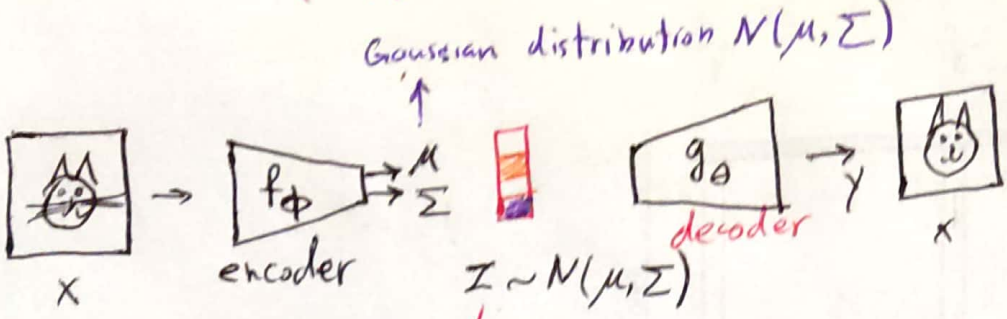
generating data

choose a random z (e.g. $z \sim N(0, I)$)

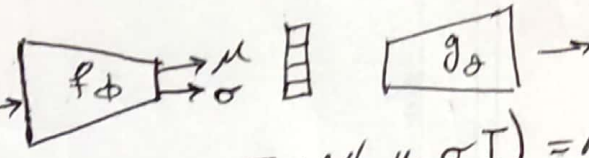
$$\hat{x} = g_\theta(z)$$

doesn't work because the distribution of z is unknown $\neq N(0, I)$

Variational Auto-Encoders



simpler model



$z \sim N(\mu, \sigma I) = N(\mu_\phi, \sigma_\phi I)$

generation

$z \sim N(0, \sigma I)$

$y = g_\theta(z)$

$\mu, \sigma = f_\phi(x)$

$z \sim N(\mu, \sigma I)$

$\hat{x} = g_\theta(z)$

$x^1, x^2, \ldots, x^m$

$g_\theta(z(\phi))$

$$C(\theta, \phi) = \sum_{i=1}^{m} \alpha \left\| g_\theta\left(z \sim N\left(\mu_\phi(x^i), \sigma_\phi(x^i) I\right)\right) - x^i \right\|^2 + \sum_{i=1}^{m} KL\left(N\left(\mu_\phi(x^i), \sigma_\phi^2(x^i) I\right) \| N(0, I)\right)$$

How to compute $\frac{\partial}{\partial \theta} C(\theta, \phi)$, $\frac{\partial}{\partial \phi} C(\theta, \phi)$?

$\mu, \sigma = f_\phi(x)$ ✓

$z \sim N(\mu, \sigma I)$?

$y = g_\theta(z)$ ✓

How to differentiate from the sampling operation?

Reparameterization trick

$$z \sim N(\mu_\phi, \sigma_\phi^2 I)$$

$$\downarrow$$

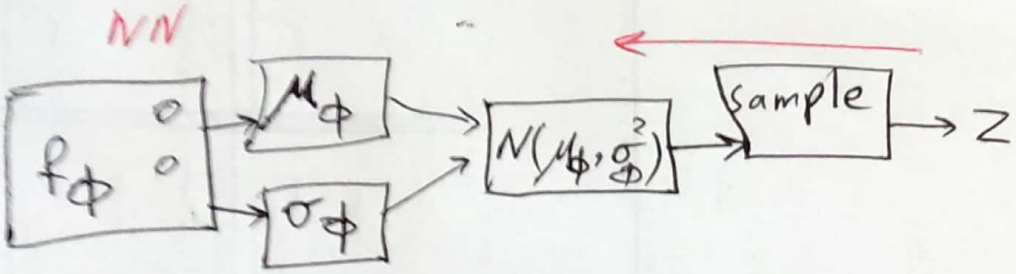
$$z = \mu_\phi + \varepsilon \sigma_\phi$$

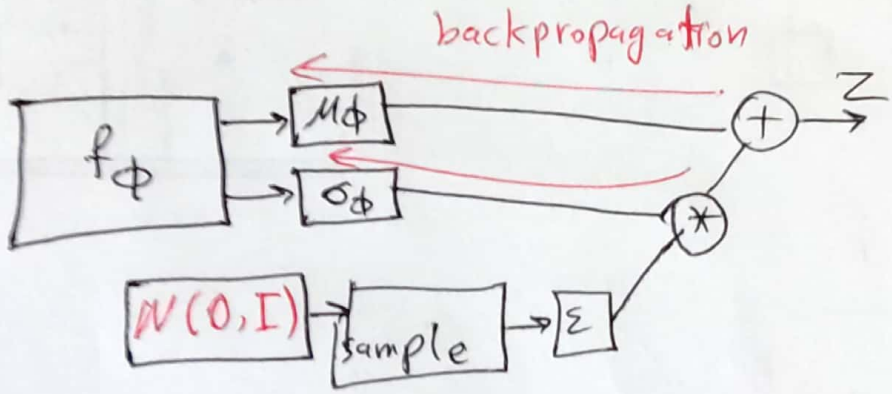
$$\varepsilon \sim N(0, I)$$

$$\varepsilon \sim N(0, I)$$

$$\sigma\varepsilon \sim N(0, \sigma^2 I)$$

$$\mu + \sigma\varepsilon \sim N(\mu, \sigma^2 I)$$





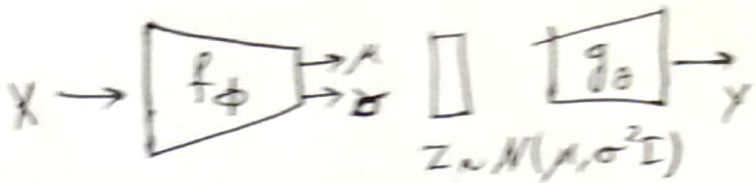
$$\varepsilon \sim N(0, I)$$

$$z = \mu_\phi + \varepsilon \sigma_\phi$$

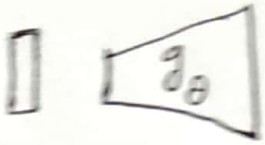
$$\frac{\partial z}{\partial \phi} = \frac{\partial \mu_\phi}{\partial \phi} + \varepsilon \frac{\partial \sigma_\phi}{\partial \phi}$$ ✓

$$\varepsilon^i \sim N(0, I)$$

$$C(\theta, \phi) = \sum_{i=1}^{m} \alpha \| g_\theta(\mu_\phi(x^i) + \varepsilon^i \sigma_\phi(x^i) I) + KL(N(\mu_\phi(x^i), \sigma_\phi^2(x^i) I) \| N(0, I))$$



generation



$z \sim N(0, I)$ $\quad Y = g_\theta(z)$ $\quad x \sim N(y, \sigma_g^2 I)$

↑ 아는 값 ↑ 아는 값

↓ 모르는 값 ↓ 모르는 값

$z \sim P(z)$ $\quad P_\theta(x|z)$

$x = N(g_\theta(z), \sigma_g^2 I)$

$P_\theta(x|z) = N(x; g_\theta(z), \sigma_g^2)$

$$P_\theta(x, z) = P_\theta(x|z)\, p(z)$$

data $x^1, x^2, \cdots, x^m$

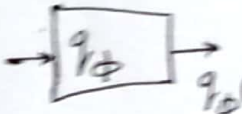
$$\max_\theta ll(\theta) = \sum_{i=1}^m \log P_\theta(x^i) = \sum_{i=1}^m \log \int P_\theta(x, z)\, dz$$

log-likelihood

$$= \sum_{i=1}^m \log \int P_\theta(x|z)\, p(z)\, dz$$

Hard to maximzied

ELBO (Amotized)



$q_\phi(z|x)$ posterior

$$\max_{\phi,\theta} L(\phi, \theta) = \sum_i \int q_\phi(z|x^i) \log \frac{P_\theta(x^i, z)}{q_\phi(z|x^i)}$$

$$= \sum_{i=1}^m \int q_\phi(z|x^i) \log \frac{P_\theta(x^i|z)\, p(z)}{q_\phi(z|x^i)}$$

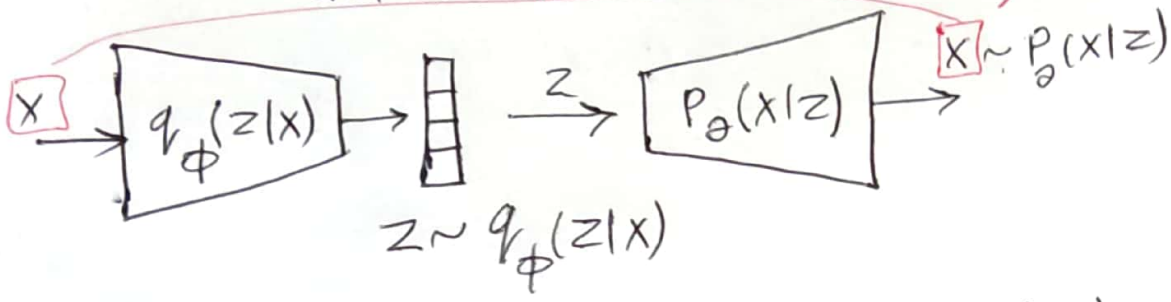
$$= \sum_{i=1}^m \int q_\phi(z|x^i) \log P_\theta(x^i|z) - \int q_\phi(z|x^i) \frac{q_\phi(z|x^i)}{p(z)}$$

$$\sum_{i=1}^m E_{q_\phi(z|x^i)}\{\log P_\theta(x^i|z)\} - KL\left(q_\phi(z|x^i) \,\|\, P(z)\right)$$

$x \rightarrow$ enc $q(x|z)$ $\quad$ dec $p(z|x)$

$$\mathcal{L}(\phi,\theta) = \sum_{i=1}^{m} E_{q_\phi(z|x^i)}\left\{\log p_\theta(x^i|z)\right\}$$

$$- \sum_{i=1}^{m} KL\left(q_\phi(z|x^i) \,\|\, p(z)\right)$$



حالت خاص $q_\phi(z|x) = \mathcal{N}\left(z;\ \mu_\phi(x^i),\ \sigma_\phi(x^i)^2 I\right)$ $\varepsilon \sim \mathcal{N}(0,$

$$z \sim q_\phi(z|x) \Rightarrow z = \mu_\phi(x^i) + \varepsilon\, \sigma_\phi(x^i)$$

$$p_\theta(x|z) = \mathcal{N}\left(x;\ g_\theta(z),\ \sigma_g I\right)$$

$$p(z) = \mathcal{N}(0, I)$$

$$\mathcal{L}(\phi,\theta) = \sum_{i=1}^{m} E_{q_\phi(z|x^i)}\left\{\log \mathcal{N}\left(x; g_\theta(z), \sigma_g I\right)\right\}$$

$$- \sum_{i=1}^{m} KL\left(\mathcal{N}(\mu_\phi(x^i), \sigma_\phi(x^i)^2 \,\|\, \mathcal{N}(0, I)\right)$$

$$= \sum_{i=1}^{m} E_{q_\phi(z|x^i)}\left\{\log \frac{1}{\sqrt{2\pi}\,\sigma_g^{d/2}} e^{\frac{\|x - g_\theta(z)\|^2}{\sigma_g^{2d}}}\right\} - KL$$

$$= \sum_{i=1}^{m} E_{q_\phi(z|x^i)}\left\{-\frac{d}{2}\log \sigma_g + \frac{\|x - g_\theta(z)\|^2}{\sigma_g^{2d}}\right\} - KL$$

one sample approximation $z^i \sim q_\phi(z|x^i)$

$$\simeq \sum \frac{\|x^i - g_\theta(z^i)\|^2}{\sigma_g^{2d}} + \sum KL$$