

# Variational Bayes with Modern Examples

Richard Xu

June 23, 2021

## 1 Maximum Likelihood Estimation

$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^n \log p_{\theta}(x_i) \quad (1)$$

as many models are defined in terms of their latent variables  $z_i$ , then we must specify  $p(x_i)$  as a marginal distribution:

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta} \sum_{i=1}^n \log \int_{z_i} p_{\theta}(x_i, z_i) \\ &= \arg \max_{\theta} \sum_{i=1}^n \log \int_{z_i} p_{\theta}(x_i | z_i) p(z_i) \end{aligned} \quad (2)$$

## 2 variational bayes

dropping index  $i$ , we want to have a good estimator of  $\log p(x|\theta)$ , we know:

$$\begin{aligned} \log p_{\theta}(x) &= \log \int_z p_{\theta}(x, z) \\ &= \log \int_z \frac{p_{\theta}(x, z|\theta)}{q_{\phi}(z|x)} q_{\phi}(z|x) \\ &= \log \left[ \mathbb{E}_{z \sim q_{\phi}(z|x)} \left( \frac{p_{\theta}(x, z|\theta)}{q_{\phi}(z|x)} \right) \right] \end{aligned} \quad (3)$$

in the above,  $\log(\mathbb{E}[\cdot])$  is not that useful, so we maximize its lower-bound, i.e., ELBO (Let's wait to see that the un-useful expression is actually the basis of IWAE)

$$\begin{aligned} &\geq \mathbb{E}_{z \sim q_{\phi}(z|x)} \left[ \log \left( \frac{p_{\theta}(x, z|\theta)}{q_{\phi}(z|x)} \right) \right] \quad \text{by Jensen's inequality} \\ &= \mathbb{E}_{z \sim q_{\phi}(z|x)} [\log(p_{\theta}(x, z|\theta))] - \mathbb{E}_{z \sim q_{\phi}(z|x)} [\log(q_{\phi}(z|x))] \\ &= \text{ELBO}(\phi) \end{aligned} \quad (4)$$

The **advantage** of ELBO is it has no “model conditional”  $p(z|x) = \frac{p(z,x)}{\int_z p(x,z)}$  (it's hard to obtain). It can be approximated by monte-carlo, using integral of  $k$  samples, where samples are from “proposal conditional”  $q_{\phi}(z|x)$

$$\begin{aligned}
\text{ELBO}(\phi) &= \mathbb{E}_{z \sim q_\phi(z|x)} \left[ \log \left( \frac{p_\theta(x, z)}{q_\phi(z|x)} \right) \right] \\
\implies \text{ELBO}_k(\phi) &= \frac{1}{k} \sum_{j=1}^k \left[ \log \left( \frac{p_\theta(x, z^j)}{q_\phi(z^j|x)} \right) \right] \\
&\text{where } z^j \sim q_\phi(z|x)
\end{aligned} \tag{5}$$

note that  $\text{ELBO}_k(\phi)$  is a  $k$  samples approximation of Monte-Carlo expectation.  
By LLN:

$$\lim_{k \rightarrow \infty} \text{ELBO}_k(\phi) = \text{ELBO}(\phi) \tag{6}$$

### 3 Evidence lower bound (ELBO)

#### 3.1 Expression ELOB

knowing:

$$\begin{aligned}
\text{ELBO}(\phi) &= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \left( \frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} p_\theta(\mathbf{x}|\mathbf{z}) \right) \right] \\
&= \int \log \left( \frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} p_\theta(\mathbf{x}|\mathbf{z}) \right) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z}
\end{aligned} \tag{7}$$

there are two main ways of expressing ELBO in literature:

- split one

$$\begin{aligned}
&= \int \log p_\theta(\mathbf{x}|\mathbf{z}) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} + \int \log \left( \frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \int \log \left( \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} \right) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}[q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})] \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}[q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})]
\end{aligned} \tag{8}$$

- split two

$$\begin{aligned}
&= \int \log p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} + \int \log \left( \frac{1}{q_\phi(\mathbf{z}|\mathbf{x})} \right) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}, \mathbf{z})] - \int \log q_\phi(\mathbf{z}|\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})] \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})]
\end{aligned} \tag{9}$$

We will document which split people are using in the following literature:

### 3.1.1 notes on the expression of ELOB

let's look at **split one** again. Since the aim of  $\text{ELBO}_{(\theta, \phi)}$  is to find alignment between  $q_\phi(\mathbf{z}|\mathbf{x})$  with the posterior  $p_\theta(\mathbf{z}|\mathbf{x})$ , then:

$$\text{ELBO}_{(\theta, \phi)} = \underbrace{\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{alignment with likelihood } p_\theta(\mathbf{x}|\mathbf{z})} + \underbrace{-\text{KL}[q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})]}_{\text{alignment with prior } p(\mathbf{z})} \quad (10)$$

therefore, we can see that  $q_\phi(\mathbf{z}|\mathbf{x})$  is the balance of the two alignments. This will be illustrated again the VAE-GAN

## 3.2 Purpose of Variational Bayes using ELBO

### 3.2.1 to approximate $p_\theta(z|x)$

using Jensen's inequality did not explicitly stating what is actually missing between  $\log p_\theta(x)$  and  $\text{ELBO}(\phi)$ , so the extract expression is:

$$\begin{aligned} \log(p_\theta(x)) &= \log(p_\theta(x, z)) - \log(p_\theta(z|x)) \\ &= \log\left(\frac{p_\theta(x, z)}{q_\phi(z|x)}\right) - \log\left(\frac{p_\theta(z|x)}{q_\phi(z|x)}\right) \\ &= \underbrace{\int q_\phi(z|x) \log\left(\frac{p_\theta(x, z)}{q_\phi(z|x)}\right) dz}_{\text{ELBO}(\phi)} + \underbrace{\left(-\int q_\phi(z|x) \log\left(\frac{p_\theta(z|x)}{q_\phi(z|x)}\right) dz\right)}_{\text{KL}(p_\theta(z|x) \| q_\phi(z|x))} \\ &= \text{ELBO}(\phi) + \text{KL}(p_\theta(z|x) \| q_\phi(z|x)) \end{aligned} \quad (11)$$

maximizing ELBO has the same effect as minimize KL, which means VB allow  $q_\phi(z|x)$  to approximate  $p_\theta(z|x)$

### 3.2.2 perform Maximum Likelihood

to perform MLE:

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta} \sum_{i=1}^n \log p_\theta(x_i) \\ &\approx \arg \max_{\theta, \phi} \sum_{i=1}^n \text{ELBO}(\phi) \quad \text{approximated by lower-bound} \\ &\approx \arg \max_{\theta, \phi} \sum_{i=1}^n \text{ELBO}_k(\phi) \quad \text{further approximated by MC integral} \\ &= \arg \max_{\theta, \phi} \sum_{i=1}^n \frac{1}{k} \sum_{j=1}^k \left[ \log \left( \frac{p_\theta(x, z^j)}{q_\phi(z^j|x)} \right) \right] \quad z^j \sim q_\phi(z^j|x) \\ &= \arg \max_{\theta, \phi} \sum_{i=1}^n \sum_{j=1}^k \left[ \log \left( \frac{p_\theta(x, z^j)}{q_\phi(z^j|x)} \right) \right] \quad z^j \sim q_\phi(z^j|x) \end{aligned} \quad (12)$$

## 4 Importance weighted auto-encoders

### 4.1 IWAE<sub>k</sub>

this section is to explain [?]

looking at Eq.(3), we know the following identity:

$$\log p_\theta(x) = \log \left[ \mathbb{E}_{z \sim q_\phi(z|x)} \left( \frac{p_\theta(x, z|\theta)}{q_\phi(z|x)} \right) \right]$$

the goal is to approximate the above; however, let us first define an expression:

$$\widehat{\text{IWAE}}_k = \log \left[ \frac{1}{k} \sum_{j=1}^k \frac{p_\theta(x|z^{(j)})p(z^{(j)})}{q_\phi(z^{(j)}|x)} \right] \quad (13)$$

Note that although  $\widehat{\text{IWAE}}_k$  looks like  $\text{ELBO}_k(\phi)$ ,  $\widehat{\text{IWAE}}_k$  was merely an expression **inside** the monte-carlo integral. It's **not** approximation to expectation. In fact, we need to “arm” it by putting this expression inside an Expectation, to make it functional:

$$\begin{aligned} \text{IWAE}_k &= \mathbb{E}_{\{z^{(j)} \sim q_\phi(z|x)\}_{j=1}^k} \left[ \widehat{\text{IWAE}}_k \right] \\ &= \mathbb{E}_{\{z^{(j)} \sim q_\phi(z|x)\}_{j=1}^k} \left[ \log \left[ \frac{1}{k} \sum_{j=1}^k \frac{p_\theta(x|z^{(j)})p(z^{(j)})}{q_\phi(z^{(j)}|x)} \right] \right] \\ &= \int_{z^{(1)}} \cdots \int_{z^{(k)}} \log \left[ \frac{1}{k} \sum_{j=1}^k \frac{p_\theta(x|z^{(j)})p(z^{(j)})}{q_\phi(z^{(j)}|x)} \right] \prod_{j=1}^k q_\phi(z^{(j)}|x) \end{aligned} \quad (14)$$

in summary,  $\text{IWAE}_k$  itself is an exact expectation of the expression  $\widehat{\text{IWAE}}_k$ . So if one is to approximate  $\text{IWAE}_k$ , one must sample, sample-set  $\{z^{(1)}, \dots, z^{(k)}\}$  multiple say  $n$  times.

Now looking at what happens when we have  $k = 1$  and  $k = \infty$ :

### 4.2 IWAE<sub>1</sub>

what if we have  $k = 1$ , by looking Eq.(20), we have:

$$\begin{aligned} \text{IWAE}_1 &= \mathbb{E}_{z^{(1)} \sim q_\phi(z|x)} \left[ \widehat{\text{IWAE}}_1 \right] \\ &= \mathbb{E}_{z^{(1)} \sim q_\phi(z|x)} \left[ \log \left[ \frac{p_\theta(x|z^{(1)})p(z^{(1)})}{q_\phi(z^{(1)}|x)} \right] \right] \\ &= \mathbb{E}_{z \sim q_\phi(z|x)} \left[ \log \left[ \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} \right] \right] \quad \text{drop index} \\ &= \text{ELBO}(\phi) \end{aligned} \quad (15)$$

### 4.3 IWAE<sub>∞</sub>

in fact, there is no need to explicitly proving IWAE<sub>∞</sub>, we can use the fact that  $\forall k$ :

$$\begin{aligned}
\text{IWAE}_k &= \mathbb{E}_{\{z^{(j)} \sim q_\phi(z|x)\}_{j=1}^k} \left[ \log \left[ \left( \frac{1}{k} \sum_{j=1}^k \frac{p_\theta(x|z^{(j)})p(z^{(j)})}{q_\phi(z^{(j)}|x)} \right) \right] \right] \\
&\leq \log \left( \mathbb{E}_{\{z^{(j)} \sim q_\phi(z|x)\}_{j=1}^k} \left[ \left( \frac{1}{k} \sum_{j=1}^k \frac{p_\theta(x|z^{(j)})p(z^{(j)})}{q_\phi(z^{(j)}|x)} \right) \right] \right) \\
&= \log \frac{1}{k} \int_{z^{(2)}} \cdots \int_{z^{(k)}} \left( \sum_{j=2}^k \frac{p_\theta(x|z^{(j)})p(z^{(j)})}{q_\phi(z^{(j)}|x)} + \underbrace{\int_{z^{(1)}} \frac{p_\theta(x|z^{(1)})p(z^{(1)})}{q_\phi(z^{(1)}|x)} q_\phi(z^{(1)}|x)}_{=p_\theta(x)} \right) \prod_{j=2}^k q_\phi(z^{(j)}|x) \\
&= \frac{kp_\theta(x)}{k} \\
&= p_\theta(x)
\end{aligned} \tag{16}$$

since the upper-bound of  $\text{IWAE}_k = p_\theta(x) \forall k$ , then, by proving section(4.4), we can deduce:

$$\text{IWAE}_\infty = p_\theta(x) \tag{17}$$

### 4.4 Tighter bound

it can be proven that:

$$\text{ELBO} = \text{IWAE}_1 \leq \text{IWAE}_2 \leq \cdots \leq \text{IWAE}_\infty = \log p_\theta(x) \tag{18}$$

#### 4.4.1 proof of why $k \geq m \implies \text{IWAE}_k \geq \text{IWAE}_m$

First, intuitively, the following is true:

$$\mathbb{E}_{I=\{j_1, \dots, j_m\}} \left[ \frac{w_{j_1} + \cdots + w_{j_m}}{m} \right] = \frac{w_1 + \cdots + w_k}{k} \tag{19}$$

What that means is that given  $m \leq k$ , you are selecting uniformly a subset of  $m$  elements from  $k$  available data. Then, instead of perform true average on  $k$ -element data, you are performing an average on the  $m$ -element subset.

In Eq.(19), it says the expectation of the “average of uniformly-drawn sub-set”, equal the value of true average. Note the above should **not** work when  $m > k$ . Also note that the original set  $\{w_1, \dots, w_k\}$  does not need to be stochastic.

Now we apply the above lemma to IWAE<sub>k</sub> equation:

$$\begin{aligned}
\text{IWAE}_k &= \mathbb{E}_{\{z^{(j)} \sim q_\phi(z|x)\}_{j=1}^k} \left[ \log \left[ \underbrace{\frac{1}{k} \sum_{j=1}^k \frac{p_\theta(x|z^{(j)})p(z^{(j)})}{q_\phi(z^{(j)}|x)}}_{\text{true average}} \right] \right] \\
&= \mathbb{E}_{\{z^{(j)} \sim q_\phi(z|x)\}_{j=1}^k} \left[ \log \left[ \underbrace{\mathbb{E}_{I=\{j_1, \dots, j_m\}} \left[ \frac{1}{m} \sum_{t=1}^m \frac{p_\theta(x|z^{(j_t)})p(z^{(j_t)})}{q_\phi(z^{(j_t)}|x)} \right]}_{\text{expectation of "average of uniformly-drawn sub-set"}} \right] \right] \\
&\geq \mathbb{E}_{\{z^{(j)} \sim q_\phi(z|x)\}_{j=1}^k} \left[ \mathbb{E}_{I=\{j_1, \dots, j_m\}} \left[ \log \left[ \frac{1}{m} \sum_{t=1}^m \frac{p_\theta(x|z^{(j_t)})p(z^{(j_t)})}{q_\phi(z^{(j_t)}|x)} \right] \right] \right] \quad \text{by Jensen's inequality}
\end{aligned} \tag{20}$$

Now looking at  $\mathbb{E}_{\{z^{(j)} \sim q_\phi(z|x)\}_{j=1}^k} [\mathbb{E}_{I=\{j_1, \dots, j_m\}} [\cdot]]$ , these two nested expectation is computed over the probability, by first selecting  $k$  i.i.d samples from  $q_\phi(z|x)$ , and then select  $m$  subset from it. (However, the above may possibly result duplicating values of  $z^{(j)}$ )

So the two integral can combine together:

$$\begin{aligned}
&= \mathbb{E}_{\{z^{(j_t)} \sim q_\phi(z|x)\}_{t=1}^m} \left[ \log \left[ \frac{1}{m} \sum_{t=1}^m \frac{p_\theta(x|z^{(j_t)})p(z^{(j_t)})}{q_\phi(z^{(j_t)}|x)} \right] \right] \\
&= \mathbb{E}_{\{z^{(j)} \sim q_\phi(z|x)\}_{j=1}^m} \left[ \log \left[ \frac{1}{m} \sum_{j=1}^m \frac{p_\theta(x|z^{(j)})p(z^{(j)})}{q_\phi(z^{(j)}|x)} \right] \right] \quad \text{drop index of } t \tag{21} \\
&= \text{IWAE}_m
\end{aligned}$$

we have proved  $k \geq m \implies \text{IWAE}_k \geq \text{IWAE}_m$

## 5 Example of Variational Inference: Normalized Flow

The first paper of VB on Normalized Flow can be found at [?]

### 5.1 Revision on Change of Variable

take Integration by substitution problem, and let

$$\mathbf{y} = f(\mathbf{x}) \implies \mathbf{x} = f^{-1}(\mathbf{y}) \tag{22}$$

let  $\mathbf{x} \in \mathbb{R}^n$  and  $\mathbf{y} \in \mathbb{R}^n$  You would like total volume to remain constant, i.e.,:

$$|\mathbf{dy} \times f(\mathbf{y})| = |\mathbf{dx} \times f(\mathbf{x})| \tag{23}$$

However, as  $f(\mathbf{y})$  and  $f(\mathbf{x})$  obviously do not equal, making  $\mathbf{dx}$  and  $\mathbf{dy}$  correspond to different infinitesimal base volume. So how are  $\mathbf{dx}$  and  $\mathbf{dy}$  related? in turns out that:

$$\underbrace{dx_1 \cdots dx_n}_{\text{corresponding infinitesimal base volume in } d\mathbf{x}} = \underbrace{\left| \det \left( \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \right|}_{\text{volume of change ratios}} \underbrace{dy_1 \cdots dy_n}_{\text{reference infinitesimal base volume in } d\mathbf{y}} \quad (24)$$

using  $d\mathbf{y}$  as the infinitesimal “reference” base volume, then the corresponding  $d\mathbf{x}$  (or  $df^{-1}(\mathbf{y})$ ) must be:

$$d\mathbf{y} \times \underbrace{\text{volume of } \underbrace{\text{instantaneous changes ratio between } \mathbf{x} \text{ and } \mathbf{y}}_{\textcircled{1}}}_{\textcircled{2}} \quad (25)$$

there are two parts to the above equation,  $\textcircled{1}$  “instantaneous changes ratio between  $\mathbf{x}$  and  $\mathbf{y}$ ” can be described by:

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \quad (26)$$

which is the Jacobian matrix w.r. to  $\mathbf{y}$ , and then, the  $\textcircled{2}$  is the volume of the parallelo-  
tope spanned by the columns of this Jacobian. This is the determinant!

$$\left| \det \left( \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) \right| \quad (27)$$

one can visualize it as following: given a reference rectangle volume (or area if it’s 2D):  $d\mathbf{y}$ , through mapping function  $f^{-1}(\mathbf{y})$ , its corresponding parallelogram  $d\mathbf{x}$ ’s volume is determined by  $\left| \det \left( \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) \right| d\mathbf{y}$ .

formally:

$$\begin{aligned} dx_1 \cdots dx_n &= \left| \det \left( \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) \right| dy_1 \cdots dy_n, \text{ or,} \\ d\mathbf{x} &= \left| \det \left( \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) \right| d\mathbf{y} \end{aligned} \quad (28)$$

### 5.1.1 apply change of variable to probability

$$\begin{aligned}
\Pr(Y \in \mathbb{S}) &= \int_{\mathbb{S}} \underbrace{p_Y(\mathbf{y})}_{\text{things inside the integral}} d\mathbf{y} \\
&= \int_{f^{-1}(\mathbb{S})} p_X(\mathbf{x}) d\mathbf{x} = \Pr(X \in f^{-1}(\mathbb{S})) \\
&= \int_{f^{-1}(\mathbb{S})} p_X(\mathbf{x}) \left| \det \left( \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) \right| d\mathbf{y} \quad \text{substitute change of variable} \\
&= \int_{\mathbb{S}} \underbrace{p_X(f^{-1}(\mathbf{y})) \left| \det \left( \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) \right|}_{\text{things inside the integral}} d\mathbf{y} \\
\Rightarrow p_Y(\mathbf{y}) &= p_X(f^{-1}(\mathbf{y})) \left| \det \left( \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) \right| \quad \text{things inside the integral} \\
&= p_X(f^{-1}(\mathbf{y})) \left| \det \left( \frac{\partial \mathbf{y}}{\partial f^{-1}(\mathbf{y})} \right) \right|^{-1} \quad \text{property of } \det(\cdot) \\
&= p_X(\mathbf{x}) \left| \det \left( \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right) \right|^{-1} \quad \text{that's the familiar expression}
\end{aligned} \tag{29}$$

### 5.1.2 one reason to have $|\det(\cdot)|$

$$\begin{aligned}
\Pr(b \leq Y \leq a) &= \int_a^b p_X(f^{-1}(\mathbf{y})) \left( \det \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) d\mathbf{y} \\
&= \int_b^a p_X(f^{-1}(\mathbf{y})) \left( -\det \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right) d\mathbf{y} \\
&= \int_a^b p_X(f^{-1}(\mathbf{y})) \left| \det \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right| d\mathbf{y}
\end{aligned} \tag{30}$$

## 5.2 apply to Normalized Flow

re-writing  $\mathbf{x} \rightarrow \mathbf{z}$ , and  $\mathbf{y} \rightarrow \mathbf{z}'$ :

$$p(\mathbf{z}') = p(\mathbf{z}) \left| \det \frac{\partial f(\mathbf{z})}{\partial \mathbf{z}} \right|^{-1} \tag{31}$$

we let:

$$\mathbf{z}_K = f_K \circ \dots \circ f_2 \circ f_1(\mathbf{z}_0) \tag{32}$$

starting backwards, and let  $\underbrace{\mathbf{z}_K}_{\mathbf{z}'} = f_K(\underbrace{\mathbf{z}_{K-1}}_{\mathbf{z}})$ :



$$\begin{aligned}
p(\mathbf{z}_K) &= \underbrace{p(\mathbf{z}_{K-1})}_{\text{}} \left| \det \frac{\partial f_K(\mathbf{z}_{K-1})}{\partial \mathbf{z}_{K-1}} \right|^{-1} \\
&= \underbrace{p(\mathbf{z}_{K-2}) \left| \det \frac{\partial f_{K-1}(\mathbf{z}_{K-2})}{\partial \mathbf{z}_{K-2}} \right|^{-1}}_{\text{}} \times \left| \det \frac{\partial f_K(\mathbf{z}_{K-1})}{\partial \mathbf{z}_{K-1}} \right|^{-1} \\
&= \vdots \\
&= p_0(\mathbf{z}_0) \left| \det \frac{\partial f_1(\mathbf{z}_0)}{\partial \mathbf{z}_0} \right|^{-1} \times \dots \times \left| \det \frac{\partial f_{K-1}(\mathbf{z}_{K-2})}{\partial \mathbf{z}_{K-2}} \right|^{-1} \times \left| \det \frac{\partial f_K(\mathbf{z}_{K-1})}{\partial \mathbf{z}_{K-1}} \right|^{-1} \\
\implies \log(p(\mathbf{z}_K)) &= \log(p_0(\mathbf{z}_0)) + \sum_{k=1}^K \log \left| \det \frac{\partial f_k(\mathbf{z}_{k-1})}{\partial \mathbf{z}_{k-1}} \right|^{-1} \\
&= \log(p_0(\mathbf{z}_0)) - \sum_{k=1}^K \log \left| \det \frac{\partial f_k(\mathbf{z}_{k-1})}{\partial \mathbf{z}_{k-1}} \right|^{-1}
\end{aligned} \tag{33}$$

### 5.2.1 Expectation

using the final equation form:

$$\log(p(\mathbf{z}_K)) = \log(p_0(\mathbf{z}_0)) - \sum_{k=1}^K \log \left| \det \frac{\partial f_k(\mathbf{z}_{k-1})}{\partial \mathbf{z}_{k-1}} \right|^{-1} \tag{34}$$

substitute it to derive expectation:

$$\begin{aligned}
\mathbb{E}_{p_K}[h(\mathbf{z})] &\equiv \mathbb{E}_{p(\mathbf{z}_K)}[h(\mathbf{z}_K)] \\
&= \int_{\mathbf{z}_K} h(\mathbf{z}_K) p(\mathbf{z}_K) d\mathbf{z}_K \\
&= \int_{\mathbf{z}_0} h(\mathbf{f}_K \circ \dots \circ \mathbf{f}_2 \circ \mathbf{f}_1(\mathbf{z}_0)) p(\mathbf{z}_0) d\mathbf{z}_0 \\
&= \mathbb{E}_{p(\mathbf{z}_0)}[h(\mathbf{f}_K \circ \dots \circ \mathbf{f}_2 \circ \mathbf{f}_1(\mathbf{z}_0))]
\end{aligned} \tag{35}$$

### 5.3 variational learning of Normalized Flow

Obviously, Normalized Flow is used in a varieties of settings. However, when it is used in ELBO, it is used in  $q_\phi(\mathbf{z})$  we replace all previous representation from  $p \rightarrow q$ , and also not explicitly writing out  $q_\phi$  for clarity:

let  $q_\phi(\mathbf{z}|\mathbf{x}) \equiv q_K(\mathbf{z}_K)$ , and substitute:

$$\log(q_\phi(\mathbf{z}_K)) = \log(q_0(\mathbf{z}_0)) - \sum_{k=1}^K \log \left| \det \frac{\partial f_k(\mathbf{z}_{k-1})}{\partial \mathbf{z}_{k-1}} \right|^{-1} \tag{36}$$

It us using **split two** of the ELBO:

$$\begin{aligned}
\text{ELBO}_{(\theta, \phi)} &= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})] \\
&= \mathbb{E}_{\mathbf{z}_K \sim q_K(\mathbf{z}_K)} [\log p_\theta(\mathbf{x}, \mathbf{z}_K) - \log q_K(\mathbf{z}_K)] && \text{using } q_K(\mathbf{z}_K) \equiv q_\phi(\mathbf{z}|\mathbf{x}) \\
&= \mathbb{E}_{\mathbf{z}_0 \sim q_0(\mathbf{z}_0)} [\log p_\theta(\mathbf{x}, \mathbf{z}_K) - \log q_K(\mathbf{z}_K)] && q_0(\mathbf{z}_0) = q_K(\mathbf{z}_K) \text{ by NF construction} \\
&= \mathbb{E}_{\mathbf{z}_0 \sim q_0(\mathbf{z}_0)} \left[ \log p_\theta(\mathbf{x}, \mathbf{z}_K) - \log(q_0(\mathbf{z}_0)) + \sum_{k=1}^K \log \left| \det \frac{\partial f_k(\mathbf{z}_{k-1})}{\partial \mathbf{z}_{k-1}} \right|^{-1} \right]
\end{aligned} \tag{37}$$

### 5.3.1 NF variational algorithm

now by keeping  $\det \frac{\partial f_k(\mathbf{z}_{k-1})}{\partial \mathbf{z}_{k-1}} = 1$ , which is **not** necessary, but convenient to make:

$$\sum_{k=1}^K \log \left| \det \frac{\partial f_k(\mathbf{z}_{k-1})}{\partial \mathbf{z}_{k-1}} \right|^{-1} = 0 \tag{38}$$

in each iteration:

$$\begin{aligned}
&\text{get mini-batch } \{\mathbf{x}\} \\
&\mathbf{z}_0 \sim q_0(\cdot|\mathbf{x}) \\
&\text{re-parameterization it as:} \\
&\epsilon \sim \mathcal{N}(0, \mathbf{I}) \\
&\mathbf{z}_0 = \mathbf{z}_{0\phi}(\mathbf{x}, \epsilon) \\
&\quad = \mu_\phi(\mathbf{x}) + \Sigma_\phi(\mathbf{x}) \times \epsilon \\
&\mathbf{z}_K \leftarrow f_K \circ \dots \circ f_2 \circ f_1(\mathbf{z}_0) \\
&\triangle \theta \propto -\nabla_\theta \text{ELBO}_{\theta, \phi}(\mathbf{x}, \mathbf{z}_K) \\
&\triangle \phi = -\nabla_\phi \text{ELBO}_{\theta, \phi}(\mathbf{x}, \mathbf{z}_K)
\end{aligned} \tag{39}$$

## 6 Variational Auto Encoder

it uses the **split one** of ELBO derivation:

$$\text{ELBO}_{(\theta, \phi)} = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \tag{40}$$

### 6.1 VAE algorithm

each iteration:

$$\begin{aligned}
& \text{get mini-batch } \{\mathbf{x}\} \\
& \mathbf{z} \sim q_\phi(\cdot|\mathbf{x}) \\
& \text{re-parameterization:} \\
& \epsilon \sim \mathcal{N}(0, \mathbf{I}) \\
& \mathbf{z} = \text{Encoder}_\phi(\mathbf{x}, \epsilon) \\
& \quad = \mu_\phi(\mathbf{x}) + \Sigma_\phi(\mathbf{x}) \times \epsilon \\
& \Delta\theta \propto -\nabla_\theta \text{ELBO}_{(\theta, \phi)}(\mathbf{x}, \mathbf{z}) \\
& \Delta\phi = -\nabla_\phi \text{ELBO}_{(\theta, \phi)}(\mathbf{x}, \mathbf{z})
\end{aligned} \tag{41}$$

### 6.1.1 evaluating $\log p_\theta(\mathbf{x}|\mathbf{z})$ through reconstruction loss

under traditional variational inference  $\log p_\theta(\mathbf{x}|\mathbf{z})$  is evaluable.

However, in the typical settings of VAE, for example where  $\mathbf{x}$  is images,  $\log p_\theta(\mathbf{x}|\mathbf{z})$  can not be evaluated.

This is of course where the backward **decoder** becomes helpful to evaluate it, i.e:

$$\hat{\mathbf{x}} = \text{Decoder}_\theta(\mathbf{z}) \tag{42}$$

therefore:

$$\begin{aligned}
p_\theta(\mathbf{x}|\mathbf{z}) &\equiv p(\mathbf{x} | \text{Decoder}_\theta(\mathbf{z})) \quad \text{by VAE} \\
&\propto \exp(-d(\mathbf{x}, \hat{\mathbf{x}} = \text{Decoder}_\theta(\mathbf{z}))) \\
&= \exp(-d(\mathbf{x}, \hat{\mathbf{x}})) \\
\implies \log p_\theta(\mathbf{x}|\mathbf{z}) &= -d(\mathbf{x}, \hat{\mathbf{x}})
\end{aligned} \tag{43}$$

making the first term just the average reconstruction loss, we may rewrite ELOB again for VAE:

$$\begin{aligned}
\text{ELBO}_{(\theta, \phi)} &= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \\
&= \mathbb{E}_{\mathbf{z} \sim \text{Encoder}_\phi(\mathbf{x})} [-d(\mathbf{x}, \text{Decoder}_\theta(\mathbf{z}))] - \text{KL}[\text{Encoder}_\phi(\mathbf{x})||p(\mathbf{z})]
\end{aligned} \tag{44}$$

## 6.2 some points to note

- $\text{Encoder}_\phi(\mathbf{x})$  is actually a re-parameterized probability density function  $q_\phi(\mathbf{z}|\mathbf{x})$ , whereas the  $\text{Decoder}_\theta(\mathbf{z})$  is only part of the probability of  $p_\theta(\mathbf{x}|\mathbf{z})$
- $p(\mathbf{z})$  are to **evaluate**  $\text{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]$ , it is not used for sampling. Therefore, in theory, one may use very complex  $p(\mathbf{z})$  form, as long as it's evaluable
- $\text{Encoder}_\phi(\mathbf{x}, \epsilon)$  is a single inference network

### 6.3 relationship with VAE-GAN

due to the claim that VAE's decoder (used for reconstruction) may not be as effective as GAN's generator ( $\text{Gen}^{\text{GAN}}$ ). Therefore, we can do the following.

We also change it to minimization instead of maximization.

By letting  $\text{Des}_l^{\text{GAN}}$  to be the  $l^{\text{th}}$  layer of Discriminator, and of course the GAN objective will be able to train  $\text{Gen}^{\text{GAN}}$ , and  $\text{Des}^{\text{GAN}}$

$$\begin{aligned}
-\text{ELBO}_{(\theta, \phi)} + \text{GAN} &= \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} [-\log p_{\theta}(\mathbf{x}|\mathbf{z})] + \text{KL}[\text{Encoder}_{\phi}(\mathbf{x}) \| p(\mathbf{z})] + \text{GAN} \\
&= \underbrace{\mathbb{E}_{\mathbf{z} \sim \text{Encoder}_{\phi}(\mathbf{x})} [-d(\mathbf{x}, \text{Decoder}_{\theta}(\mathbf{z}))]}_{\text{replace}} + \underbrace{\text{KL}[q_{\phi}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})]}_{\text{keep alignment with prior}} + \text{GAN} \\
&= E_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \left[ -\log p_{\theta} \left( \text{Des}_l^{\text{GAN}}(\mathbf{x}) \mid \text{Des}_l^{\text{GAN}}(\text{Decoder}_{\theta}(\mathbf{z})) \right) \right] + \text{KL}[\text{Encoder}_{\phi}(\mathbf{x}) \| p(\mathbf{z})] + \text{GAN} \\
&= E_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \left[ -\log \mathcal{N} \left( \text{Des}_l^{\text{GAN}}(\mathbf{x}) ; \text{Des}_l^{\text{GAN}}(\text{Decoder}_{\theta}(\mathbf{z})) \right) \right] + \text{KL}[\text{Encoder}_{\phi}(\mathbf{x}) \| p(\mathbf{z})] + \text{GAN}
\end{aligned} \tag{45}$$

#### 6.3.1 notes on VAE-GAN

there could be many different implementation to the above. for example:

- one may let

$$E_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \left[ -\log \mathcal{N} \left( \text{Des}_l^{\text{GAN}}(\mathbf{x}) ; \text{Des}_l^{\text{GAN}}(\text{Decoder}_{\theta}(\mathbf{z})) \right) \right] \text{ to also update } \text{Des}^{\text{GAN}} \text{ parameters}$$

- one may replace:

$$\mathcal{N} \left( \text{Des}_l^{\text{GAN}}(\mathbf{x}) ; \text{Des}_l^{\text{GAN}}(\text{Decoder}_{\theta}(\mathbf{z})) \right) \rightarrow \mathcal{N} \left( \text{Des}_l^{\text{GAN}}(\mathbf{x}) ; \text{Des}_l^{\text{GAN}}(\text{Gen}^{\text{GAN}}) \right) \tag{46}$$

### 6.4 KL between two Gaussian distributions

Last piece of puzzle is that, VAE objective function requires to compute KL between two Gaussians, let's have a look at their forms:

#### 6.4.1 generalized for to compute $\text{KL}(\mathcal{N}(\mu_1, \Sigma_1) \| \mathcal{N}(\mu_2, \Sigma_2))$

$$\begin{aligned}
\text{KL} &= \int_x \left[ \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \right] \times p(x) dx \\
&= \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2} \text{tr} \left\{ \mathbb{E}[(x - \mu_1)(x - \mu_1)^T] \Sigma_1^{-1} \right\} + \frac{1}{2} \mathbb{E}[(x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2)] \\
&= \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2} \text{tr} \{I_d\} + \frac{1}{2} (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2) + \frac{1}{2} \text{tr} \{ \Sigma_2^{-1} \Sigma_1 \} \\
&= \frac{1}{2} \left[ \log \frac{|\Sigma_2|}{|\Sigma_1|} - d + \text{tr} \{ \Sigma_2^{-1} \Sigma_1 \} + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right]
\end{aligned} \tag{47}$$

substitute  $\bar{\mu}_1 = [\mu_1, \dots, \mu_K]^T$  and  $\Sigma_1 = \text{diag}(\sigma_1, \dots, \sigma_K)$ ,  $\mu_2 = \mathbf{0}$  and  $\Sigma_2 = \mathbf{I}$ :

$$\begin{aligned}
\text{KL} &= \frac{1}{2} \left( \text{tr}(\Sigma_1) + \bar{\mu}_1^T \bar{\mu}_1 - K - \log \det(\Sigma_1) \right) \\
&= \frac{1}{2} \left( \sum_k \sigma_k^2 + \sum_k \mu_k^2 - \sum_k 1 - \log \prod_k \sigma_k^2 \right) \\
&= \frac{1}{2} \sum_k (\sigma_k^2 + \mu_k^2 - 1 - \log \sigma_k^2)
\end{aligned} \tag{48}$$

**6.4.2** when  $p(x_1, x_2) = p(x_1)p(x_2)$  and  $q(x_1, x_2) = q(x_1)q(x_2)$  (1)

$$\begin{aligned}
\text{KL}(p, q) &= - \left( \int p(x_1) \log q(x_1) dx_1 - \int p(x_1) \log p(x_1) dx_1 \right) \\
&\implies \text{KL}(p(x_1)p(x_2) \| q(x_1)q(x_2)) \\
&= - \left( \int_{x_1} \int_{x_2} p(x_1)p(x_2) [\log q(x_1) + \log q(x_2)] dx_1 - p(x_1)p(x_2) [\log p(x_1) + \log p(x_2)] dx_1 \right) \\
&= - \left( \int_{x_1} \int_{x_2} [p(x_1)p(x_2) \log q(x_1) + p(x_1)p(x_2) \log q(x_2) - p(x_1)p(x_2) \log p(x_1) - p(x_1)p(x_2) \log p(x_2)] dx_1 \right) \\
&= - \left( \int_{x_1} \int_{x_2} p(x_1)p(x_2) \log q(x_1) + \int_{x_1} \int_{x_2} p(x_1)p(x_2) \log q(x_2) - \int_{x_1} \int_{x_2} p(x_1)p(x_2) \log p(x_1) - \int_{x_1} \int_{x_2} p(x_1)p(x_2) \log p(x_2) \right) \\
&= - \left( \int_{x_1} p(x_1) \log q(x_1) \int_{x_2} p(x_2) + \int_{x_1} p(x_1) \int_{x_2} p(x_2) \log q(x_2) - \int_{x_1} p(x_1) \log p(x_1) \int_{x_2} p(x_2) - \int_{x_1} p(x_1) \int_{x_2} p(x_2) \log p(x_2) \right) \\
&= - \left( \int_{x_1} p(x_1) \log q(x_1) + \int_{x_2} p(x_2) \log q(x_2) - \int_{x_1} p(x_1) \log p(x_1) - \int_{x_2} p(x_2) \log p(x_2) \right) \\
&= - \left( \int_{x_1} p(x_1) \log q(x_1) - \int_{x_1} p(x_1) \log p(x_1) \right) - \left( \int_{x_2} p(x_2) \log q(x_2) - \int_{x_2} p(x_2) \log p(x_2) \right) \\
&= \text{KL}(p(x_1) \| q(x_1)) + \text{KL}(p(x_2) \| q(x_2))
\end{aligned} \tag{49}$$

therefore,

$$\begin{aligned}
&\text{KL}(p(x_1)p(x_2) \| q(x_1)q(x_2)) = \text{KL}(p(x_1) \| q(x_1)) + \text{KL}(p(x_2) \| q(x_2)) \\
&\implies \text{KL} \left( \prod_k p(x_k) \| \prod_k q(x_k) \right) = \sum_{i=1}^k \text{KL}(p(x_i) \| q(x_i))
\end{aligned} \tag{50}$$

**6.4.3** when  $p(x_1, x_2) = p(x_1)p(x_2)$  and  $q(x_1, x_2) = q(x_1)q(x_2)$  (2)

let  $p(x) = \mathcal{N}(\mu_p, \sigma_p)$  and  $q(x) = \mathcal{N}(\mu_q, \sigma_q)$ :

$$\begin{aligned}
\text{KL}(p, q) &= - \int p(x) \log q(x) dx + \int p(x) \log p(x) dx \\
&= \frac{1}{2} \log(2\pi\sigma_q^2) + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}(1 + \log 2\pi\sigma_p^2) \\
&= \log \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2} \\
&= \log \sigma_q - \log \sigma_p + \frac{\sigma_p^2}{2\sigma_q^2} + \frac{(\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}
\end{aligned} \tag{51}$$

let  $p(x) = \mathcal{N}(\mu, \sigma)$  and  $q(x) = \mathcal{N}(0, 1)$ :

$$\begin{aligned}
\text{KL}(p, q) &= \frac{\sigma^2}{2} + \frac{\mu^2}{2} - \frac{1}{2} - \log \sigma \\
&= \frac{1}{2} \left[ \frac{\sigma^2}{2} + \frac{\mu^2}{2} - \frac{1}{2} - \log \sigma^2 \right]
\end{aligned} \tag{52}$$

moving into  $k$  dimensions, and apply  $\text{KL}\left(\prod_k p(x_k) \parallel \prod_k q(x_k)\right) = \sum_{i=1}^k \text{KL}(p(x_i) \parallel q(x_i))$ :

$$\text{KL}\left(\prod_k p(x_k) \parallel \prod_k q(x_k)\right) = \frac{1}{2} \sum_k \left[ \frac{\sigma^2}{2} + \frac{\mu^2}{2} - \frac{1}{2} - \log \sigma^2 \right] \tag{53}$$

## 7 Adversarial Variational Bayes

This section is to explain [?]  
it uses **split one** of ELBO:

$$\begin{aligned}
\text{ELBO}_{(\theta, \phi)} &= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{x}|\mathbf{z}) \right] - \text{KL} \left[ q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}) \right] \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{x}|\mathbf{z}) - \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} \right] \\
&= \max_{\psi} \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{x}|\mathbf{z}) - T_\psi(\mathbf{x}, \mathbf{z}) \right] \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{x}|\mathbf{z}) - T_\psi^*(\mathbf{x}, \mathbf{z}) \right]
\end{aligned} \tag{54}$$

the paper ignores structure of  $\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}$  and train to obtain  $T_\psi^*(\mathbf{x}, \mathbf{z})$  complete separate network.

in VAE, one needs to assume how to **evaluate**  $q_\phi(\mathbf{z}|\mathbf{x})$  to be some distribution, in AVB, we treat it as black-box inference model, we only need to know how to sample from  $q_\phi(\mathbf{z}|\mathbf{x})$

## 7.1 how do you obtain $T_\psi^*(\mathbf{x}, \mathbf{z})$

we use the following objective function:

$$T_\psi^*(\mathbf{x}, \mathbf{z}) = \max_{\psi} \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \sigma(T_\psi(\mathbf{x}, \mathbf{z})) \right] + \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{p(\mathbf{z})} \left[ \log(1 - \sigma(T_\psi(\mathbf{x}, \mathbf{z}))) \right] \quad (55)$$

logistic regression to differentiate  $(\mathbf{x}, \mathbf{z})$  between  $\underbrace{p(\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x})}_{\text{real}}$  and  $\underbrace{p(\mathbf{x})p(\mathbf{z})}_{\text{fake}}$

note that we didn't use  $p(\mathbf{x}, \mathbf{z})$  but instead  $p(\mathbf{x})$  and  $p(\mathbf{z})$

### 7.1.1 why does this objective work?

we must prove the following, let:

$$T_\psi^*(\mathbf{x}, \mathbf{z}) = \max_{\psi} \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \sigma(T_\psi(\mathbf{x}, \mathbf{z})) \right] + \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{p(\mathbf{z})} \left[ \log(1 - \sigma(T_\psi(\mathbf{x}, \mathbf{z}))) \right] \quad (56)$$

this implies:

$$\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{x}|\mathbf{z}) - T_\psi^*(\mathbf{x}, \mathbf{z}) \right] = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{x}|\mathbf{z}) \right] - \text{KL} \left[ q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}) \right] \quad (57)$$

i.e., after  $\max_{\psi}$ , we get our original ELBO back.

### 7.1.2 proof is similarity to GAN's optimum $D^*(\mathbf{x})$

look at GAN after fix  $G$  and optimize  $D$ : (see my GAN notes):

$$\begin{aligned} & \max_D \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{x} \sim p_g^\theta(\mathbf{x})} [\log(1 - D(\mathbf{x}))] \\ \implies D^*(x) &= \frac{p_r(x)}{p_r(x) + p_g^\theta(x)} \end{aligned} \quad (58)$$

compare it with Eq.(55) and to look at pattern, the best  $\sigma(T^*(\mathbf{x}, \mathbf{z}))$  should occur when:

$$\begin{aligned} \sigma(T^*(\mathbf{x}, \mathbf{z})) &= \frac{p(\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x}) + p(\mathbf{x})p(\mathbf{z})} \\ &= \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x}) + p(\mathbf{z})} \\ &= \frac{q}{q + p} \quad \text{for simple notation} \end{aligned} \quad (59)$$

$$\begin{aligned}
&\implies \frac{1}{1 + \exp(-T^*)} = \frac{q}{q + p} \quad \text{definition of } \sigma \\
&\implies q + p = q(1 + \exp(-T^*)) \\
&\implies p = q \exp(-T^*) \\
&\implies \log \frac{p}{q} = -T^* \\
&\implies T_\psi^* = \log(q_\phi(\mathbf{z}|\mathbf{x})) - \log p(\mathbf{z}) \quad \text{substitute back in}
\end{aligned} \tag{60}$$

in summary, by calculating:

$$\begin{aligned}
T_\psi^*(\mathbf{x}, \mathbf{z}) &= \max_{\psi} \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \sigma(T(\mathbf{x}, \mathbf{z})) \right] + \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{p(\mathbf{z})} \left[ \log(1 - \sigma(T(\mathbf{x}, \mathbf{z}))) \right] \\
&\implies \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{x}|\mathbf{z}) - T_\psi^*(\mathbf{x}, \mathbf{z}) \right] = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{x}|\mathbf{z}) \right] - \text{KL} \left[ q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}) \right]
\end{aligned} \tag{61}$$

## 7.2 Overall objective

$$\max_{\theta} \max_{\phi} \max_{\psi} \left[ \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \sigma(T_\psi(\mathbf{x}, \mathbf{z})) \right] + \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{p(\mathbf{z})} \left[ \log(1 - \sigma(T_\psi(\mathbf{x}, \mathbf{z}))) \right] \right] \tag{62}$$