Variational Bayes with Modern Examples

Richard Xu

April 27, 2022

1 Maximum Likelihood Estimation

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

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

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

2 variational bayes

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

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

in the above, $\log(\mathbb{E}[.])$ 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)

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

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" $q_{\phi}(z|x)$

2.1 monte-carlo approximation

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

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

$$
\lim_{k\to\infty} \mathrm{ELBO}_k(\phi) = \mathrm{ELBO}(\phi) \tag{6}
$$

3 Evidence lower bound (ELBO)

3.1 Expression ELOB

knowing:

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

there are two main ways of expressing ELBO in literature:

- split **one**

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

the advantage is that we can express it in terms of the KL. Let's look at **split one**, we can view the aim of $\mathrm{ELBO}_{(\theta,\phi)}$ to be finding alignment between $q_\phi(\mathbf{z}|\mathbf{x})$ with the posterior $p_\theta(\mathbf{z}|\mathbf{x})$:

$$
\mathrm{ELBO}_{(\theta,\phi)} = \underbrace{\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathrm{x}|\mathrm{z})\big]}_{\text{alignment with likelihood} p_\theta(\mathbf{x}|\mathbf{z})} + \underbrace{-\mathrm{KL}\big[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\big]}_{\text{alignment with prior } p(z)} \tag{9}
$$

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

- split **two**

$$
\begin{aligned}
&= \int \log p_\theta(\mathrm{x}|\mathrm{z}) p(\mathbf{z}) q_\phi(\mathbf{z}|\mathbf{x}) \mathrm{d}\mathbf{z} + \int \log \left(\frac{1}{q_\phi(\mathbf{z}|\mathbf{x})} \right) q_\phi(\mathbf{z}|\mathbf{x}) \mathrm{d}\mathbf{z} \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \big[\log p_\theta(\mathrm{x}, \mathrm{z}) \big] - \int \log q_\phi(\mathbf{z}|\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x}) \mathrm{d}\mathbf{z} \\
&= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \big[\log p_\theta(\mathrm{x}, \mathrm{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}) \big]
\end{aligned} \tag{10}
$$

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

3.2 Purpose of Variational Bayes using ELBO

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

We already stated that $p(z|x) = \frac{p(z,x)}{\int_z p(x,z)}$ is difficult to compute. Jensen's inequality did not explicitly stating what is actually missing between $\log p_\theta(x)$ and ELBO(ϕ), so the extract expression is:

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

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:

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

4 Importance weighted auto-encoders

4.1 IWAE$_k$

this section is to explain [1]. 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:

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

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. Itself is a random variable, it's **not** an approximation to expectation. In fact, we need to "arm" it by putting this expression inside an Expectation, to make it functional:

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

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

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

4.2 IWAE$_1$

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

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

4.3 IWAE$_\infty$

in fact, there is no need to explicitly proving IWAE$_\infty$, we can use the fact that $\forall k$:

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

since the upper-bound of 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 $k \geq m \implies$ IWAE$_k \geq$ IWAE$_m$

First, intuitively, the following is true:

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

In words, the *"average of a uniformly generated sub-set equal the average of a full-set"*.

More formally, it 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 $\{w_1, \ldots w_k\}$ does not need to be stochastic.

Now we apply the above lemma to IWAE$_k$ equation:

$$
\begin{aligned}
\text{IWAE}_k &= \mathbb{E}_{\left\{z^{(j)} \sim q_\phi(z|x)\right\}_{j=1}^k} \left[\log \left[\underbrace{\frac{1}{k} \sum_{j=1}^{k} \frac{p_\theta(x|z^{(j)})p(z^{(j)})}{q_\phi(z^{(j)}|x)}}_{\text{true average}} \right] \right] \\
&= \mathbb{E}_{\left\{z^{(j)} \sim q_\phi(z|x)\right\}_{j=1}^k} \left[\log \left[\underbrace{\mathbb{E}_{I=\{j_1,\dots,j_m\}} \left[\frac{1}{m} \sum_{t=1}^{m} \frac{p_\theta(x|z^{(j_t)})p(z^{(j_t)})}{q_\phi(z^{(j_t)}|x)} \right]}_{\text{expectation of ``average of uniformly-drawn sub-set''}} \right] \right] \quad \text{apply Eq.(19)} \\
&\geq \mathbb{E}_{\left\{z^{(j)} \sim q_\phi(z|x)\right\}_{j=1}^k} \left[\mathbb{E}_{I=\{j_1,\dots,j_m\}} \left[\log \left[\frac{1}{m} \sum_{t=1}^{m} \frac{p_\theta(x|z^{(j_t)})p(z^{(j_t)})}{q_\phi(z^{(j_t)}|x)} \right] \right] \right] \quad \text{by Jensen's inequality}
\end{aligned} \tag{20}
$$

Now looking at $\mathbb{E}_{\left\{z^{(j)} \sim q_\phi(z|x)\right\}_{j=1}^k} \left[\mathbb{E}_{I=\{j_1,\dots,j_m\}}[.]\right]$, 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:

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

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

5 Variational Auto Encoder

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

$$\text{ELBO}_{(\theta,\phi)} = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathrm{x}|\mathrm{z})\big] - \text{KL}\big[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\big] \tag{22}$$

note that if we use **split one**:

$$\text{ELBO}_{(\theta,\phi)} = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathrm{x},\mathrm{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\big] \tag{23}$$

although it's the same thing, but we cannot have a nice KL interpretation.

5.1 VAE algorithm

during each iteration of gradient descend, the gradient is computed as:

$$\begin{aligned}
&\text{get mini-batch } \{\mathbf{x}\} \\
&\mathbf{z} \sim q_\phi(\cdot|\mathbf{x}) \\
&\qquad \text{re-parameterization:} \\
&\qquad\ \ \epsilon \sim \mathcal{N}(0, \mathbf{I}) \\
&\qquad\ \ \mathbf{z} = \text{Encoder}_\phi(\mathbf{x}, \epsilon) \\
&\qquad\qquad = \mu_\phi(\mathbf{x}) + \Sigma_\phi(\mathbf{x}) \times \epsilon \\
&\triangle\theta \propto -\nabla_\theta \text{ELBO}_{(\theta,\phi)}(\mathbf{x}, \mathbf{z}) \\
&\triangle\phi = -\nabla_\phi \text{ELBO}_{(\theta,\phi)}(\mathbf{x}, \mathbf{z})
\end{aligned} \tag{24}$$

5.1.1 evaluating $\log p_\theta(\mathrm{x}|\mathrm{z})$ through reconstruction loss

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

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

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

$$\hat{\mathbf{x}} = \text{Decoder}_\theta(\mathbf{z}) \tag{25}$$

therefore:

$$\begin{aligned}
p_\theta(\mathrm{x}|\mathrm{z}) &\equiv p\big(\mathbf{x} \mid \text{Decoder}_\theta(\mathbf{z})\big) \qquad \text{by VAE} \\
&\propto \exp\big(-d(\mathbf{x},\ \hat{\mathbf{x}} = \text{Decoder}_\theta(\mathbf{z}))\big) \\
&= \exp\big(-d(\mathbf{x},\ \hat{\mathbf{x}})\big) \\
\implies \log p_\theta(\mathbf{x}|\mathbf{z}) &= -d(\mathbf{x},\ \hat{\mathbf{x}})
\end{aligned} \tag{26}$$

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

$$
\begin{aligned}
\mathrm{ELBO}_{(\theta,\phi)} &= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathrm{x}|\mathrm{z})\big] - \mathrm{KL}\big[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\big] \\
&= \mathbb{E}_{\mathbf{z}\sim \mathrm{Encoder}_\phi(\mathbf{x})}\big[-d\big(\mathrm{x}, \mathrm{Decoder}_\theta(\mathrm{z})\big)\big] - \mathrm{KL}\big[\mathrm{Encoder}_\phi(\mathbf{x})\|p(\mathbf{z})\big]
\end{aligned} \tag{27}
$$

5.2 some points to note

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

5.3 relationship with VAE-GAN

5.3.1 why VAE generate blur image

Due to the claim that VAE's decoder (used for reconstruction) may not be as effective as GAN's generator ($\mathrm{Gen}^{\mathrm{GAN}}$). A popular explanation of why VAE may generate blur image: one explanation is that if reconstruction loss was $d\big(\mathrm{x}, \mathrm{Decoder}_\theta(\mathrm{z})\big)$, and imagine $\mathrm{Decoder}_\theta(\mathbf{z})$ is a blur version of $\mathbf{x}$, then, their VAE-reconstruction loss is in fact small (They can be "content-wise" similar, but "style-wise" different - think about an image and its Gaussian smooth version can have small L_2 loss). GAN on the other hand has no individual reconstructions. Therefore, it is looking for global distribution similarity (style loss)

5.3.2 VAE-GAN loss

Therefore, we can do the following, and we also change the objective to minimization instead of maximization.

By letting $\mathrm{Des}_l^{\mathrm{GAN}}$ to be the l^{th} layer of Discriminator (therefore, $\mathrm{Des}_l^{\mathrm{GAN}} \in \mathbb{R}^{m_l}$ whereas $\mathrm{Des}^{\mathrm{GAN}} \in (0, \dots 1)$). Of course the GAN objective will be able to train $\mathrm{Gen}^{\mathrm{GAN}}$, and $\mathrm{Des}^{\mathrm{GAN}}$

$$
\begin{aligned}
-\mathrm{ELBO}_{(\theta,\phi)} + \mathrm{GAN} &= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\big[-\log p_\theta(\mathrm{x}|\mathrm{z})\big] + \mathrm{KL}\big[\mathrm{Encoder}_\phi(\mathbf{x})\|p(\mathbf{z})\big] + \mathrm{GAN} \\
&= \underbrace{\mathbb{E}_{\mathbf{z}\sim \mathrm{Encoder}_\phi(\mathbf{x})}\big[-d\big(\mathrm{x}, \mathrm{Decoder}_\theta(\mathrm{z})\big)\big]}_{\text{replace}} + \underbrace{\mathrm{KL}\big[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\big]}_{\text{keep alignment with prior}} + \mathrm{GAN} \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Big[-\log p_\theta\Big(\mathrm{Des}_l^{\mathrm{GAN}}(\mathrm{x}) \mid \mathrm{Des}_l^{\mathrm{GAN}}(\mathrm{Decoder}_\theta(\mathrm{z}))\Big)\Big] + \mathrm{KL}\big[\mathrm{Encoder}_\phi(\mathbf{x})\|p(\mathbf{z})\big] + \mathrm{GAN} \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Big[-\log \mathcal{N}\Big(\mathrm{Des}_l^{\mathrm{GAN}}(\mathrm{x})\ ;\ \mathrm{Des}_l^{\mathrm{GAN}}(\mathrm{Decoder}_\theta(\mathrm{z}))\Big)\Big] + \mathrm{KL}\big[\mathrm{Encoder}_\phi(\mathbf{x})\|p(\mathbf{z})\big] + \mathrm{GAN}
\end{aligned} \tag{28}
$$

Whereas in VAE-GAN-reconstruction loss $d\Big(\mathrm{Des}_l^{\mathrm{GAN}}(\mathrm{x})\ , \mathrm{Des}_l^{\mathrm{GAN}}(\mathrm{Decoder}_\theta(\mathrm{z}))\Big)$ needs to ensure they are "style-wise" similar features. Putting both, real data $\mathbf{x}$ and $\tilde{x}$ from decoder $\tilde{x} = \mathrm{Decoder}_\theta(\mathbf{z})$ into Discriminator.

5.3.3 notes on VAE-GAN

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

- if one were to replace $\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Big[-\log\mathcal{N}\Big(\mathrm{Des}^{\mathrm{GAN}}_l(\mathbf{x})\ ;\ \mathrm{Des}^{\mathrm{GAN}}_l(\mathrm{Decoder}_\theta(\mathbf{z}))\Big)\Big]$ to also update $\mathrm{Des}^{\mathrm{GAN}}$ with

$$\mathcal{N}\Big(\mathrm{Des}^{\mathrm{GAN}}_l(\mathbf{x})\ ;\ \mathrm{Des}^{\mathrm{GAN}}_l(\mathrm{Decoder}_\theta(\mathbf{z}))\Big)\rightarrow\mathcal{N}\Big(\mathrm{Des}^{\mathrm{GAN}}_l(\mathbf{x})\ ;\ \mathrm{Des}^{\mathrm{GAN}}_l(\mathrm{Gen}^{\mathrm{GAN}}(\mathbf{z}))\Big) \tag{29}$$

  it makes no sense, as we are not learning decoder parameter.

5.4 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:

5.4.1 generallized for to compute KL$\big(\mathcal{N}(\mu_1,\Sigma_1)\|\mathcal{N}(\mu_2,\Sigma_2)\big)$

$$\begin{aligned}
\mathrm{KL} &= \int_x\left[\frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right]\times p(x)\mathrm{d}x\\
&= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}-\frac{1}{2}\mathrm{tr}\ \left\{\mathbb{E}[(x-\mu_1)(x-\mu_1)^T]\ \Sigma_1^{-1}\right\}+\frac{1}{2}\mathbb{E}[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)]\\
&= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}-\frac{1}{2}\mathrm{tr}\ \{I_d\}+\frac{1}{2}(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)+\frac{1}{2}\mathrm{tr}\{\Sigma_2^{-1}\Sigma_1\}\\
&= \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|}-d+\mathrm{tr}\{\Sigma_2^{-1}\Sigma_1\}+(\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right]
\end{aligned} \tag{30}$$

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

$$\begin{aligned}
\mathrm{KL} &= \frac{1}{2}\ \Big(\mathrm{tr}(\Sigma_1)+\bar{\mu}_1^T\bar{\mu}_1-K-\log\ \det(\Sigma_1)\Big)\\
&= \frac{1}{2}\ \left(\sum_k\sigma_k^2+\sum_k\mu_k^2-\sum_k 1-\log\prod_k\sigma_k^2\right)\\
&= \frac{1}{2}\ \sum_k\big(\sigma_k^2+\mu_k^2-1-\log\ \sigma_k^2\big)
\end{aligned} \tag{31}$$

5.4.2 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)

$$
\begin{aligned}
&\mathrm{KL}(p, q) = -\left(\int p(x_1)\log q(x_1)\mathrm{d}x_1 - \int p(x_1)\log p(x_1)\mathrm{d}x_1\right) \\
&\implies \mathrm{KL}(p(x_1)p(x_2)\|q(x_1)q(x_2)) \\
&= -\left(\int_{x_1}\int_{x_2} p(x_1)p(x_2)\big[\log q(x_1) + \log q(x_2)\big]\mathrm{d}x_1 - p(x_1)p(x_2)\big[\log p(x_1) + \log p(x_2)\big]\mathrm{d}x_1\right) \\
&= -\left(\int_{x_1}\int_{x_2} \big[p(x_1)p(x_2)\log q(x_1) + p(x_1)p(x_2)\log q(x_2) - p(x_1)p(x_2)\log p(x_1) - p(x_1)p(x_2)\log p(x_2)\big]\mathrm{d}x_1\right) \\
&= -\left(\int_{x_1}\int_{x_2} p(x_1)p(x_2)\log q(x_1) + \int_{x_1}\int_{x_2} p(x_1)p(x_2)\log q(x_2) - \int_{x_1}\int_{x_2} p(x_1)p(x_2)\log p(x_1) - \int_{x_1}\int_{x_2} p(x_1)p(x_2)\log \right. \\
&= -\left(\int_{x_1} p(x_1)\log q(x_1)\int_{x_2} p(x_2) + \int_{x_1} p(x_1)\int_{x_2} p(x_2)\log q(x_2) - \int_{x_1} p(x_1)\log p(x_1)\int_{x_2} p(x_2) - \int_{x_1} p(x_1)\int_{x_2} p(x_2)\,\mathrm{l} \right. \\
&= -\left(\int_{x_1} p(x_1)\log q(x_1) + \int_{x_2} p(x_2)\log q(x_2) - \int_{x_1} p(x_1)\log p(x_1) - \int_{x_2} p(x_2)\log p(x_2)\right) \\
&= -\left(\int_{x_1} p(x_1)\log q(x_1) - \int_{x_1} p(x_1)\log p(x_1)\right) - \left(\int_{x_2} p(x_2)\log q(x_2) - \int_{x_2} p(x_2)\log p(x_2)\right) \\
&= \mathrm{KL}(p(x_1)\|q(x_1)) + \mathrm{KL}(p(x_2)\|q(x_2))
\end{aligned}
\tag{32}
$$

therefore,

$$
\begin{aligned}
\mathrm{KL}(p(x_1)p(x_2)\|q(x_1)q(x_2)) &= \mathrm{KL}(p(x_1)\|q(x_1)) + \mathrm{KL}(p(x_2)\|q(x_2)) \\
\implies \mathrm{KL}\left(\prod_k p(x_k)\|\prod_k q(x_k)\right) &= \sum_{i=1}^{k}\mathrm{KL}(p(x_i)\|q(x_i))
\end{aligned}
\tag{33}
$$

5.4.3 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 $q(x) = \mathcal{N}(\mu_q, \sigma_q)$:

$$
\begin{aligned}
\mathrm{KL}(p, q) &= -\int p(x)\log q(x)\mathrm{d}x + \int p(x)\log p(x)\mathrm{d}x \\
&= \frac{1}{2}\log(2\pi\sigma_q^2) + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}(1 + \log 2\pi\sigma_p^2) \\
&= \log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2} \\
&= \log\sigma_q - \log\sigma_p + \frac{\sigma_p^2}{2\sigma_q^2} + \frac{(\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}
\end{aligned}
\tag{34}
$$

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

$$
\begin{aligned}
\mathrm{KL}(p, q) &= \frac{\sigma^2}{2} + \frac{\mu^2}{2} - \frac{1}{2} - \log\sigma \\
&= \frac{1}{2}\left[\frac{\sigma^2}{2} + \frac{\mu^2}{2} - \frac{1}{2} - \log\sigma^2\right]
\end{aligned}
\tag{35}
$$

moving into k dimensions, and apply $\mathrm{KL}\bigg(\prod_k p(x_k) \| \prod_k q(x_k)\bigg) = \sum_{i=1}^{k} \mathrm{KL}(p(x_i) \| q(x_i))$:

$$\mathrm{KL}\bigg(\prod_k p(x_k) \| \prod_k q(x_k)\bigg) = \frac{1}{2}\sum_k \left[\frac{\sigma^2}{2} + \frac{\mu^2}{2} - \frac{1}{2} - \log \sigma^2\right] \tag{36}$$

6 Gaussian Mixture Model variational inference

6.1 model

This was from the paper [2]:

$$\begin{aligned}
\mathbf{w} &\sim \mathcal{N}(0, \mathbf{I}) \\
\mathbf{z} &\sim \text{Mult}(\boldsymbol{\pi}) \\
\mathbf{x}|\mathbf{z}, \mathbf{w} &\sim \prod_{k=1} \mathcal{N}\big(\boldsymbol{\mu}_{z_k}(\mathbf{w}; \beta), \text{diag}\big(\boldsymbol{\sigma}^2_{z_k}(\mathbf{w}; \beta)\big)\big)^{z_k} \qquad \text{where } z_k \in \{0, 1\} \\
\mathbf{y}|\mathbf{x} &\sim \mathcal{N}\big(\boldsymbol{\mu}(\mathbf{x}; \theta), \text{diag}\big(\boldsymbol{\sigma}^2(\mathbf{w}; \theta)\big)\big)
\end{aligned} \tag{37}$$

Looking at the graphical model, it may also work if $\mathbf{w}$ is set to a fixed hyper-parameter. Basically, $\mathbf{x}$ is the $\mathbf{z}$ in the conventional VAE. The following is the relationship between the conventional and new representation:

$$p(\mathbf{z}) \longrightarrow p(\mathbf{w})p(\mathbf{z})p_\beta(\mathbf{x}|\mathbf{w}, \mathbf{z}) \tag{38}$$

note that in conventional VAE, $p(\mathbf{z})$ has no parameters. The decoder part is the similar:

$$p(\mathbf{x}|\mathbf{z}) \longrightarrow p_\theta(\mathbf{y}|\mathbf{x}) \tag{39}$$

6.2 choose appropriate $q(\cdot)$

the conventional $q(\mathbf{z}|\mathbf{x})$ becomes:

$$q(\mathbf{x}, \mathbf{w}, \mathbf{z}|\mathbf{y}) = \prod_i q_{\phi_x}(\mathbf{x}_i|\mathbf{y}_i) q_{\phi_w}(\mathbf{w}_i|\mathbf{y}_i) p_\beta(\mathbf{z}_i|\mathbf{x}_i, \mathbf{w}_i) \tag{40}$$

the key to the paper is that the prior $p_\beta(\mathbf{z}_i|\mathbf{x}_i, \mathbf{w}_i)$ also becomes part of $q(\cdot)$:

$$\begin{aligned}
\text{ELBO} &= \mathbb{E}_q\Big[\frac{p_{\beta,\theta}(\mathbf{y}, \mathbf{x}, \mathbf{w}, \mathbf{z})}{q(\mathbf{x}, \mathbf{w}, \mathbf{z}|\mathbf{y})}\Big] \\
&= \mathbb{E}_q\Big[\log\Big(\frac{p_\theta(\mathbf{y}|\mathbf{x})p(\mathbf{w})p(\mathbf{z})p_\beta(\mathbf{x}|\mathbf{w}, \mathbf{z})}{q_{\phi_x}(\mathbf{x}|\mathbf{y})q_{\phi_w}(\mathbf{w}|\mathbf{y})p_\beta(\mathbf{z}|\mathbf{x}, \mathbf{w})}\Big)\Big] \\
&= \mathbb{E}_q\Big[\log\Big(p_\theta(\mathbf{y}|\mathbf{x})\Big)\Big] + \mathbb{E}_q\Big[\log\Big(\frac{p_\beta(\mathbf{x}|\mathbf{w}, \mathbf{z})}{q_{\phi_x}(\mathbf{x}|\mathbf{y})}\Big)\Big] + \mathbb{E}_q\Big[\log\Big(\frac{p(\mathbf{w})}{q_{\phi_w}(\mathbf{w}|\mathbf{y})}\Big)\Big] + \mathbb{E}_q\Big[\log\Big(\frac{p(\mathbf{z})}{p_\beta(\mathbf{z}|\mathbf{x}, \mathbf{w})}\Big)\Big]
\end{aligned} \tag{41}$$

Note that the $q(\cdot)$ used in the expectation is $q(\mathbf{x}, \mathbf{w}, \mathbf{z}|\mathbf{y}) = q_{\phi_x}(\mathbf{x}|\mathbf{y})q_{\phi_w}(\mathbf{w}|\mathbf{y})p_\beta(\mathbf{z}|\mathbf{x}, \mathbf{w})$, therefore we can omit terms contain variables do **not** appear inside the expectation. Also we rewrite the expectation into separate terms that participate towards KL:

$$
\begin{aligned}
\text{ELBO} &= \mathbb{E}_{q(\mathbf{x}|\mathbf{y})}\Big[\log\Big(p_\theta(\mathbf{y}|\mathbf{x})\Big)\Big] + \mathbb{E}_{q_{\phi_w}(\mathbf{w}|\mathbf{y})p_\beta(\mathbf{z}|\mathbf{x},\mathbf{w})}\mathbb{E}_{q_{\phi_x}(\mathbf{x}|\mathbf{y})}\Big[\log\Big(\frac{p_\beta(\mathbf{x}|\mathbf{w},\mathbf{z})}{q_{\phi_x}(\mathbf{x}|\mathbf{y})}\Big)\Big] \\
&\quad + \mathbb{E}_{q_{\phi_w}(\mathbf{w}|\mathbf{y})}\Big[\log\Big(\frac{p(\mathbf{w})}{q_{\phi_w}(\mathbf{w}|\mathbf{y})}\Big)\Big] + \mathbb{E}_{q_{\phi_x}(\mathbf{x}|\mathbf{y})q_{\phi_w}(\mathbf{w}|\mathbf{y})}\mathbb{E}_{p_\beta(\mathbf{z}|\mathbf{x},\mathbf{w})}\Big[\log\Big(\frac{p(\mathbf{z})}{p_\beta(\mathbf{z}|\mathbf{x},\mathbf{w})}\Big)\Big] \\
&= \mathbb{E}_{q(\mathbf{x}|\mathbf{y})}\Big[\log\Big(p_\theta(\mathbf{y}|\mathbf{x})\Big)\Big] - \mathbb{E}_{q_{\phi_w}(\mathbf{w}|\mathbf{y})p_\beta(\mathbf{z}|\mathbf{x},\mathbf{w})}\Big[\text{KL}(q_{\phi_x}(\mathbf{x}|\mathbf{y})\|p_\beta(\mathbf{x}|\mathbf{w},\mathbf{z}))\Big] \\
&\quad - \text{KL}[q_{\phi_w}(\mathbf{w}|\mathbf{y})\|p(\mathbf{w})] - \mathbb{E}_{q_{\phi_x}(\mathbf{x}|\mathbf{y})q_{\phi_w}(\mathbf{w}|\mathbf{y})}\big[\text{KL}(p_\beta(\mathbf{z}|\mathbf{x},\mathbf{w})\|p(\mathbf{z}))\big]
\end{aligned}
\tag{42}
$$

$$
\begin{aligned}
p_\beta(z_{i,j}=1|\mathbf{x},\mathbf{w}) &= \frac{p(z_{i,j}=1)p(\mathbf{x}|z_{i,j}=1,\mathbf{w})}{\sum_{k=1}^K p(z_{i,k}=1)p(\mathbf{x}|z_{i,k}=1,\mathbf{w})} \\
&= \frac{\boldsymbol{\pi}_i\mathcal{N}p(\mathbf{x}|z_j=1,\mathbf{w})}{\sum_{k=1}^K p(z_j=k)p(\mathbf{x}|z_j=k,\mathbf{w})}
\end{aligned}
\tag{43}
$$

7 Stick-breaking VAE

this comes from the paper [3]:

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

where: $\boldsymbol{\pi} = \{\pi_1, \pi_2, \dots \pi_\infty\}$:

$$
\begin{aligned}
& v_k \sim \text{Beta}(1,\alpha) \\
& \pi_k = v_k\prod_{l=1}^{k-1}(1-v_l) \\
& \theta_k \sim H \\
& G_0 = \sum_{k=1}^{\infty}\pi_k\delta_{\theta_k}
\end{aligned}
\tag{45}
$$

7.1 re-parameterization

unlike the VAE algorithm from Eq.(24) where one can re-parameterize:

$$
\begin{aligned}
& \text{re-parameterization:} \\
& \quad \epsilon \sim \mathcal{N}(0,\mathbf{I}) \\
& \quad \mathbf{z} = \text{Encoder}_\phi(\mathbf{x},\epsilon) \\
& \quad\quad = \mu_\phi(\mathbf{x}) + \Sigma_\phi(\mathbf{x})\times\epsilon
\end{aligned}
\tag{46}
$$

one may not do the same if $q_\phi(v_k|\mathbf{x})$ has a beta distribution, i.e., beta distribution does not generate non-central re-parameterization. Therefore we need to have:

$$q(v) \equiv \text{Kumaraswamy}(v; a, b) = abv^{a-1}(1 - v^a)^{b-1} \tag{47}$$

since one can re-parameterize it through the inverse of CDF:

$$v = (1 - u^{\frac{1}{b}})^{\frac{1}{a}} \qquad u \sim \text{Uniform}(0, 1) \tag{48}$$

8 Adversarial Variational Bayes

This section is to explain [4]

it uses **split one** of ELBO:

$$\begin{aligned}
\text{ELBO}_{(\theta,\phi)} &= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathrm{x}|\mathrm{z})\Big] - \text{KL}\Big[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\Big] \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Bigg[\log p_\theta(\mathrm{x}|\mathrm{z}) - \log\frac{q_\phi(z|x)}{p(z)}\Bigg] \\
&= \max_\psi \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Bigg[\log p_\theta(\mathrm{x}|\mathrm{z}) - T_\psi(\mathbf{x}, \mathbf{z})\Bigg] \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Bigg[\log p_\theta(\mathrm{x}|\mathrm{z}) - T^*_\psi(\mathbf{x}, \mathbf{z})\Bigg]
\end{aligned} \tag{49}$$

the paper ignores structure of $\log\frac{q_\phi(z|x)}{p(z)}$ 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 $q_\phi(\mathbf{z}|\mathbf{x})$

8.1 how do you obtain $T^*_\psi(\mathbf{x}, \mathbf{z})$

we use the following objective function:

$$T^*_\psi(\mathbf{x}, \mathbf{z}) = \max_\psi \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log\sigma(T_\psi(\mathbf{x}, \mathbf{z}))\Big] + \mathbb{E}_{p(\mathrm{x})}\mathbb{E}_{p(\mathrm{z})}\Big[\log(1 - \sigma(T_\psi(\mathbf{x}, \mathbf{z})))\Big] \tag{50}$$

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

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

8.1.1 why does this objective work?

we must prove the following lemma:

Lemma 1 *by defining* $T^*_\psi(\mathbf{x},\mathbf{z})$ *to be:*

$$T^*_\psi(\mathbf{x},\mathbf{z}) = \max_\psi \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log\sigma(T(\mathbf{x},\mathbf{z}))\Big] + \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{p(\mathbf{z})}\Big[\log(1-\sigma(T(\mathbf{x},\mathbf{z})))\Big] \tag{51}$$

we then have:

$$\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathrm{x}|\mathrm{z}) - T^*_\psi(\mathbf{x},\mathbf{z})\Big] = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathrm{x}|\mathrm{z})\Big] - KL\Big[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\Big] \tag{52}$$

i.e., after $\max_\psi$, we get our original ELBO back. Consequentially, we have the following overall objective:

8.1.2 overall objective

$$\begin{aligned}
&\max_\theta \max_\phi \mathrm{ELBO}_{(\theta,\phi)} \\
&= \max_\theta \max_\phi \big[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathrm{x}|\mathrm{z}) - T^*_\psi(\mathbf{x},\mathbf{z})\big]\big] \\
&= \max_\theta \max_\phi \Big[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathrm{x}|\mathrm{z}) - \max_\psi \big[\mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log\sigma(T_\psi(\mathbf{x},\mathbf{z}))\big] + \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{p(\mathbf{z})}\big[\log(1-\sigma(T_\psi(\mathbf{x},\mathbf{z})))\big]\big]\Big]\Big] \\
&= \max_\theta \max_\phi \min_\psi \Big[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathrm{x}|\mathrm{z})\big] - \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log\sigma(T_\psi(\mathbf{x},\mathbf{z}))\big] + \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{p(\mathbf{z})}\big[\log(1-\sigma(T_\psi(\mathbf{x},\mathbf{z})))\big]\Big]
\end{aligned} \tag{53}$$

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

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

$$\begin{aligned}
&\max_D \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{r}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{x}\sim p^\theta_{\mathrm{g}}(\mathbf{x})}[\log(1-D(\mathbf{x})] \\
\implies\ & D^*(x) = \frac{p_{\mathrm{r}}(x)}{p_{\mathrm{r}}(x) + p^\theta_{\mathrm{g}}(x)}
\end{aligned} \tag{54}$$

compare it with Eq.(50) and to look at pattern, the best $\sigma(T^*(\mathbf{x},\mathbf{z}))$ should occur when:

$$\begin{aligned}
\sigma(T^*(\mathbf{x},\mathbf{z})) &= \frac{p(\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x}) + p(\mathrm{x})p(\mathrm{z})} \\
&= \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x}) + p(\mathrm{z})} \\
&= \frac{q}{q+p} \quad \text{for simple notation}
\end{aligned} \tag{55}$$

$$\begin{aligned}
\implies \frac{1}{1+\exp(-T^*)} &= \frac{q}{q+p} \qquad \text{definition of } \sigma \\
\implies q+p &= q(1+\exp(-T^*)) \\
\implies p &= q\exp(-T^*) \\
\implies \log\frac{p}{q} &= -T^* \\
\implies T^*_\psi &= \log(q_\phi(\mathbf{z}|\mathbf{x})) - \log p(\mathrm{z})
\end{aligned} \tag{56}$$

in summary, by calculating:

$$
\begin{aligned}
T_\psi^*(\mathbf{x}, \mathbf{z}) &= \max_\psi \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log \sigma(T(\mathbf{x}, \mathbf{z}))\Big] + \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{p(\mathbf{z})}\Big[\log(1 - \sigma(T(\mathbf{x}, \mathbf{z})))\Big] \\
\implies & \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathrm{x}|\mathrm{z}) - T_\psi^*(\mathbf{x}, \mathbf{z})\Big] \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathrm{x}|\mathrm{z}) - \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\Big] \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathrm{x}|\mathrm{z})\big] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))
\end{aligned}
\tag{57}
$$