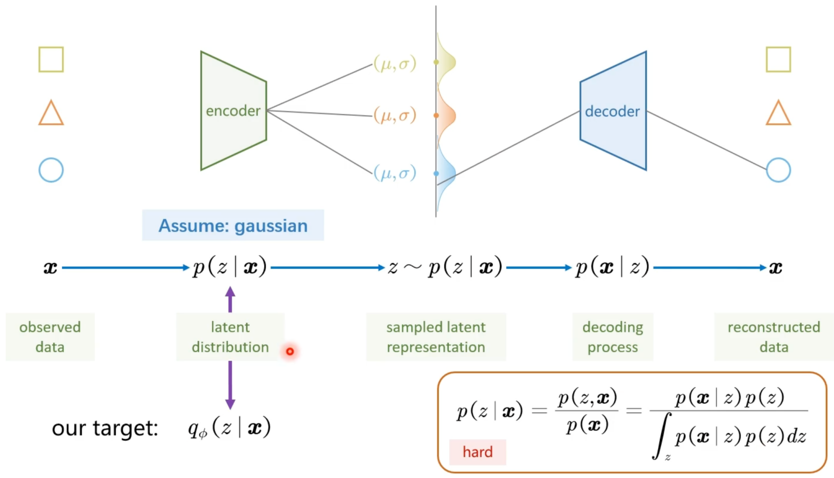
Variational Auto Encoder (VAE)

Section 1: Understand VAE



对于模型，做出如下假设

$$x \xrightarrow{Enc,p(z|\boldsymbol{x})} z \sim p(z|x) \xrightarrow{Dec,p(x|z)} x$$

而我们需要拟合 $p(z \mid x)$ 和 $p(x \mid z)$。我们拟合的概率分布如下

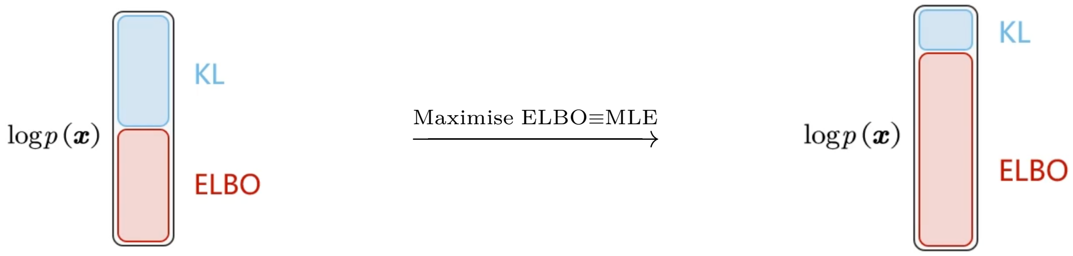
$$q_\phi(\, z \mid x\,) \approx p(\, z \mid x\,) \qquad\qquad p_\theta(\, x \mid z\,) \approx p(\, x \mid z\,)$$

根据 Naïve Bayes：

$$p(\, z \mid x\,) = \frac{p(x,z)}{p(x)} = \frac{p(x \mid z)p(z)}{p(x)} = \frac{p(x \mid z)p(z)}{\underbrace{\int_z p(x \mid z)p(z)dz}_{\text{Intractable}}}$$

ELBO: Evidence Lower Bound

$$\begin{aligned}
KL[q_\phi(\, z \mid x\,) \parallel p(\, z \mid x\,)] &= \int_z q_\phi(\, z \mid x\,) \log \frac{q_\phi(\, z \mid x\,)}{p(\, z \mid x\,)} dz = \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{q_\phi(\, z \mid x\,)}{p(\, z \mid x\,)} \right] \\
&= \int_z q_\phi(\, z \mid x\,) \log \frac{q_\phi(\, z \mid x\,)p(x)}{p(x,z)} dz \\
&= \int_z q_\phi(\, z \mid x\,) \log \frac{q_\phi(\, z \mid x\,)}{p(x,z)} dz + \int_z q_\phi(\, z \mid x\,) \log p(x) \\
&= \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{q_\phi(\, z \mid x\,)}{p(x,z)} \right] + \log p(x) \\
\underbrace{\log p(x)}_{\text{MLE}} &= \underbrace{\mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p(x,z)}{q_\phi(\, z \mid x\,)} \right]}_{\text{ELBO } \mathcal{L}} + \underbrace{KL[q_\phi(\, z \mid x\,) \parallel p(\, z \mid x\,)]}_{\epsilon}
\end{aligned}$$



因此优化目标从 MLE 转换为优化 ELBO

$$\phi^* = \underset{\phi}{\operatorname{argmax}}\, \mathbb{E}_{q_\phi(z\,|\,x)}\left[\log \frac{p(x,z)}{q_\phi(\,z \mid x\,)}\right]$$

$$\begin{aligned}
\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(\,z \mid x\,)}\right] &= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x \mid z)p(z)}{q_\phi(\,z \mid x\,)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] + \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(z)}{q_\phi(\,z \mid x\,)}\right] \\
&= \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)]}_{\text{Reconstruction Term}} - \underbrace{KL[q_\phi(\,z \mid x\,) \parallel p(z)]}_{\text{Prior Matching Term}}
\end{aligned}$$

Reconstruction Term: 可以看作是通过 Encoder 获得的 term $z \sim q_\phi(\,z \mid x\,))$，之后对这个 z 进行重建出 $\hat{x} \sim p_\theta(\,x \mid z\,)$，并对其进行 MLE。因此使其最大化，也就是使输入 x 和输出 $\hat{x}$ 的重建差值最小。也就是最小化一个 L2：

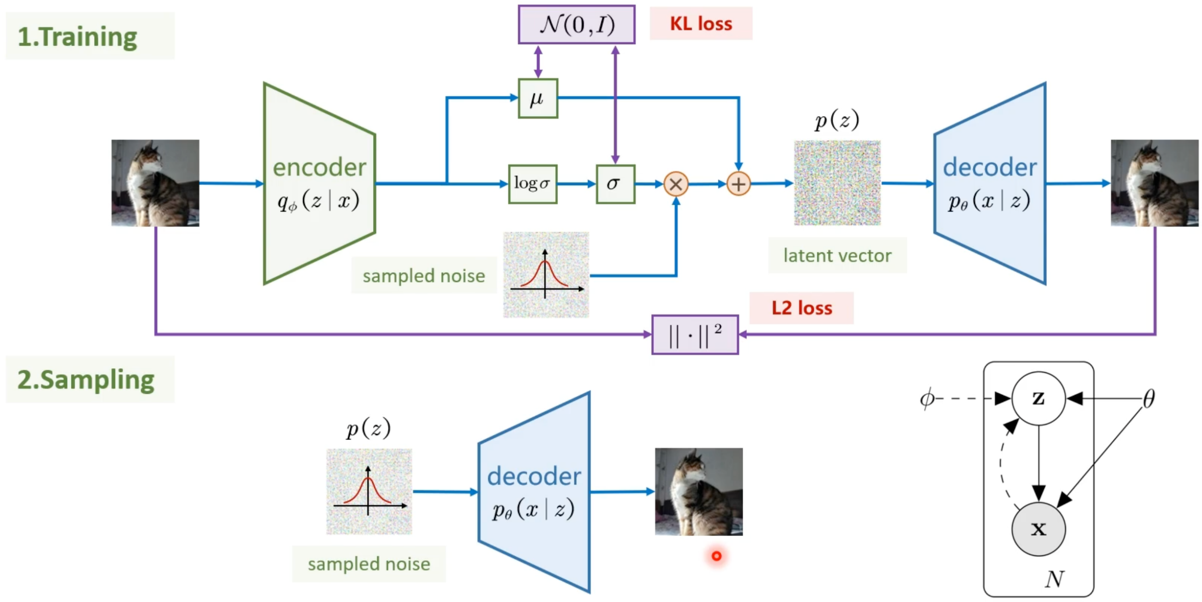
$$\text{Minimise } \|x - \hat{x}\|^2$$

Prior Matching Term：我们这里假设了 Latent 是一个 Gaussian Distribution，也就是 $p(\,z \mid x\,) \to \mathcal{N}(0, I)$。（如果不考虑方差为 1，则会退化成 AE）。

$$\begin{aligned}
p(z) &= \int_x p(\,z \mid x\,)p(x)dx \\
&= \int_x \mathcal{N}(0,I)p(x)dx \\
&= \mathcal{N}(0,I)\int_x p(x)dx \\
&= \mathcal{N}(0,I)
\end{aligned}$$

而我们也定义了 $q_\phi(\,z \mid x\,) = \mathcal{N}(z; \mu_\phi, \sigma_\phi^2 I)$，因此 PMT 可以重写为：

$$\begin{aligned}
KL[q_\phi(\,z \mid x\,) \parallel p(z)] &= KL[\mathcal{N}(z; \mu_\phi, \sigma_\phi^2 I) \parallel \mathcal{N}(0, I)] \\
&= \frac{1}{2}(\mu_\phi + \sigma_\phi^2 - \log \sigma_\phi^2 - 1)
\end{aligned}$$

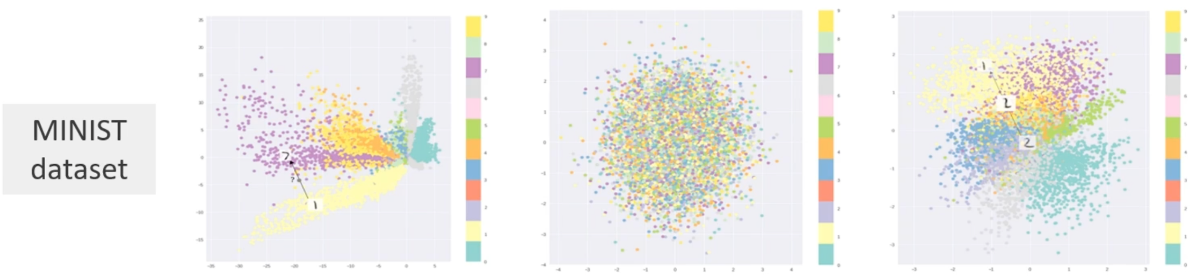


max min

$$\mathbb{E}_{q_\phi(z|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|z)] - KL(q_\phi(z|\boldsymbol{x})||p(z))$$

reconstruction term prior matching term

$$||\boldsymbol{x}-\hat{\boldsymbol{x}}||^2 \qquad \frac{1}{2}(\mu_\phi + \sigma_\phi^2 - \log\sigma_\phi^2 - 1)$$



only reconstruction loss only KL divergence combination

Section 2: Mathematics behind VAE

Mathematics Background

KL Divergence:

$$\begin{aligned}
KL[p(x) \parallel q(x)] &= \int_x p(x) \log \frac{p(x)}{q(x)} dx \\
&= -\int_x p(x) \log \frac{q(x)}{p(x)} dx \\
&\geq -\log \int_x p(x) \frac{q(x)}{p(x)} dx = 0
\end{aligned}$$

MLE:

$$\begin{aligned}
\theta^* &= \underset{\theta}{\operatorname{argmin}}\, KL[p_{data}(x) \parallel p_\theta(x)] \\
KL[p_{data}(x) \parallel p_\theta(x)] &= \underbrace{\mathbb{E}_{p_{data}(x)}[\log p_{data}(x)]}_{\text{Constant}} - \underbrace{\mathbb{E}_{p_{data}(x)}[\log p_\theta(x)]}_{\text{Depend on } \theta} \\
\theta^* &= \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{p_{data}(x)}[\log p_\theta(x)] \\
&= \underset{\theta}{\operatorname{argmax}} \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \log p_\theta(x)
\end{aligned}$$

Variational Inference

即我们需要拟合 Decoder 的生成项的分布（Latent Variable Model, LVM），也就是

$$p_\theta(x) = \int_z p(z) p_\theta(x \mid z) dx$$

这里做出的假设是

$$p(z) = \mathcal{N}(z; \mathbf{0}, \mathbf{I}) \qquad p_\theta(x \mid z) = \mathcal{N}(x; G_\theta(z), \sigma^2 I)$$

引入 z 的概率分布 $z \sim q(z)$

$$\begin{aligned}
\log p_\theta(x) &= \log \int_z p(z) p_\theta(\, x \mid z\,) dx \\
&= \log \int_z q(z) \frac{p(z) p_\theta(\, x \mid z\,)}{q(z)}\ dx \\
&\geq \int_z q(z) \log \frac{p(z) p_\theta(\, x \mid z\,)}{q(z)}\ dx \\
&= \int_z q(z) \log p_\theta(\, x \mid z\,)\ dx - \int_z q(z) \log \frac{q(z)}{p(z)}\ dx
\end{aligned}$$

$$= \underbrace{\mathbb{E}_{q(z)}[\log p_\theta(\, x \mid z \,)] - KL[q(z) \parallel p(z)]}_{\triangleq \mathcal{L}(x,q,\theta)}$$

$q(z)$ 的选择非常重要，注意到

$$\begin{aligned}
\log p_\theta(x) - KL[q(z) \parallel p_\theta(\, z \mid x \,)] &= \log p_\theta(x) - \mathbb{E}_{q(z)}\left[\log \frac{q(z)}{p_\theta(\, z \mid x \,)}\right] \\
&= \log p_\theta(x) + \mathbb{E}_{q(z)}\left[\log \frac{p_\theta(\, z \mid x \,)p(z)}{q(z)p_\theta(z)}\right] \\
&= \log p_\theta(x) + \mathbb{E}_{q(z)}\left[\log \frac{p_\theta(\, z \mid x \,)}{p_\theta(z)}\right] + \mathbb{E}_{q(z)}\left[\log \frac{p(z)}{q(z)}\right] \\
&= \mathbb{E}_{q(z)}[\log p_\theta(x)] + \mathbb{E}_{q(z)}\left[\log \frac{p_\theta(\, z \mid x \,)}{p_\theta(z)}\right] \\
- KL[q(z) \parallel p(z)] \\
&= \underbrace{\mathbb{E}_{q(z)}[\log p_\theta(\, x \mid z \,)] - KL[q(z) \parallel p(z)]}_{\mathcal{L}(x,q,\theta)} \\
\log p_\theta(x) &= \underbrace{\mathcal{L}(x,q,\theta)}_{\text{ELBO}} + \underbrace{KL[q(z) \parallel p_\theta(\, z \mid x \,)]}_{\epsilon}
\end{aligned}$$

也就是对于 MLE 的近似误差是 $KL[q(z) \parallel p_\theta(\, z \mid x \,)]$。

Variational Autoencoder

我们定义 Encoder $q(z \mid x)$为：

$$q_\phi(\, z \mid x \,) = \mathcal{N}\left(z; \mu_\phi(x), diag\left(\sigma_\phi^2(x)\right)\right) \qquad \mu_\phi, \log \sigma_\phi = \text{NN}_\phi(x)$$

$$\log\ : \mathbb{R}^+ \mapsto \mathbb{R}, \text{NN}: \mathbb{R}^d \mapsto \mathbb{R} \vdash \sigma_\phi \in \mathbb{R}^+$$

因此可以定义优化问题：

$$\begin{aligned}
\phi^*, \theta^* &= \underset{\phi,\theta}{\operatorname{argmax}}\, \mathbb{E}_{p_{data}(x)}[\log p_\theta(x)] \\
&= \underset{\phi,\theta}{\operatorname{argmax}}\, \mathbb{E}_{p_{data}(x)} \underbrace{\left[\mathbb{E}_{q_\phi(z \mid x)}[\log p_\theta(\, x \mid z \,)] - KL\big[q_\phi(z \mid x) \parallel p(z)\big]\right]}_{\mathcal{L}(x,\phi,\theta)}
\end{aligned}$$

$$\mathcal{L}(x,\phi,\theta) = \underbrace{\mathbb{E}_{q_\phi(z \mid x)}[\log p_\theta(\, x \mid z \,)]}_{\text{Reconstruction Error}} - \underbrace{KL\big[q_\phi(z \mid x) \parallel p(z)\big]}_{\text{Regularizer}}$$

可以对 Regularizer 进行解析优化（Analytic Form）

给定 $q_\phi(z \mid x)$ 和 $p(z)$ 是 factorised Gaussian distributions，则有解析式：

单变量情况分析：

$$
\begin{aligned}
& KL[q_\phi(z \mid x) \parallel p(z)] \\
&= KL[\mathcal{N}(z;\mu_\phi,\sigma_\phi^2) \parallel \mathcal{N}(z;0,I)] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{\frac{1}{\sqrt{2\pi\sigma_\phi^2}}\exp\left[-\frac{1}{2\sigma_\phi^2}(z-\mu_\phi)^2\right]}{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}z^2\right]}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{\frac{1}{\sigma_\phi}\exp\left[-\frac{1}{2\sigma_\phi^2}(z-\mu_\phi)^2\right]}{\exp\left[-\frac{1}{2}z^2\right]}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log\frac{1}{\sigma_\phi}\right] + \mathbb{E}_{q_\phi(z|x)}\left[\log\exp\left[-\frac{1}{2\sigma_\phi^2}(z-\mu_\phi)^2\right]\right] - \mathbb{E}_{q_\phi(z \mid x)}\left[\log\exp\left[-\frac{1}{2}z^2\right]\right] \\
&= -\mathbb{E}_{q_\phi(z|x)}[\log\sigma_\phi] + \mathbb{E}_{q_\phi(z|x)}\left[-\frac{1}{2\sigma_\phi^2}\underbrace{(z-\mu_\phi)^2}_{\sigma_\phi^2}\right] + \mathbb{E}_{q_\phi(z|x)}\left[\frac{1}{2}z^2\right] \\
&= -\mathbb{E}_{q_\phi(z|x)}[\log\sigma_\phi] - \frac{1}{2} + \frac{1}{2}\mathbb{E}_{q_\phi(z|x)}[(z-\mu_\phi)^2 + 2z\mu_\phi - \mu_\phi^2] \\
&= -\log\sigma_\phi - \frac{1}{2} + \frac{1}{2}\sigma_\phi^2 + \underbrace{\mathbb{E}_{q_\phi(z|x)}\underbrace{[z\mu_\phi]}_{z_i=\mu_i}}_{\mu_\phi^2} - \frac{1}{2}\mu_\phi^2 \\
&= -\log\sigma_\phi - \frac{1}{2} + \frac{1}{2}(\sigma_\phi^2 + \mu_\phi^2) \\
&= \frac{1}{2}[(\sigma_\phi^2 + \mu_\phi^2) - 2\log\sigma_\phi - 1]
\end{aligned}
$$

多维变量情况：

$$
\begin{aligned}
KL[q(z) \parallel p(z)] &= \mathbb{E}_{q(z)}\left[\log\frac{\prod_{i=1}^d q(z_i)}{\prod_{i=1}^d p(z_i)}\right] = \mathbb{E}_{q(z)}\left[\sum_{i=1}^d \log\frac{q(z_i)}{p(z_i)}\right] \\
&= \sum_{i=1}^d \mathbb{E}_{q(z_i)}\left[\sum_{i=1}^d \log\frac{q(z_i)}{p(z_i)}\right] = \sum_{i=1}^d KL[q(z_i) \parallel p(z_i)]
\end{aligned}
$$

代入单变量解析，可得

$$
KL[q_\phi(z \mid x) \parallel p(z)] = \frac{1}{2}\left(\|\mu_\phi(x)\|_2^2 + \|\sigma_\phi(x)\|_2^2 - 2\underbrace{\langle \log\sigma_\phi(x), 1\rangle}_{\sum_{i=1}^d \log\sigma_i} - d\right)
$$

会发现 Regularizer 只和 ϕ 有关。至此 Regulariser 优化完毕。

$$\mathcal{L}(x,\phi,\theta) = \underbrace{\mathbb{E}_{q_\phi(z\,|\,x)}[\log p_\theta(\,x \mid z\,)]}_{\text{Reconstruction Error}} - \underbrace{KL[q_\phi(z \mid x) \parallel p(z)]}_{\text{Regularizer}}$$

重建损失仍然 intractable，因为存在数学期望。使用 MC Estimate (Monte Carlo)：

$$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\,x \mid z\,)] \approx \log p_\theta(\,x \mid z\,) \quad z \sim p_\phi(z \mid x)$$

因此梯度可以被表示为

$$\nabla_\theta \mathcal{L}(x,\phi,\theta) \approx \nabla \log p_\theta(\,x \mid z\,) \qquad z \sim p_\phi(z \mid x)$$

$$\nabla_\phi \mathcal{L}(x,\phi,\theta) \approx \nabla_\phi \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\,x \mid z\,)] - \nabla_\phi KL[q_\phi(z \mid x) \parallel p(z)]$$

可以发现第一项仍为期望，也需要进行 MC 估计（重参数技巧 Reparameterisation Trick）

Reparameterisation Trick

$$\nabla_\phi \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\,x \mid z\,)]$$

如直接对其 MC 估计时，其需要采样 $z_i \sim q_\phi(z \mid z)$，这个采样依赖参数 ϕ，因此使用 RT 去去除对参数 ϕ 的依赖。

可以注意到：

$$z \sim p_\phi(\,z \mid x\,) \iff z = \mu_\phi + \sigma_\phi \odot \epsilon \quad \epsilon \sim \mathcal{N}(\epsilon; 0, I)$$

记 $\pi(\epsilon) := \mathcal{N}(\epsilon; 0, I)$, $T_\phi(x,\epsilon) = \mu_\phi + \sigma_\phi \odot \epsilon$，即 $z \sim T_\phi(x,\epsilon)$

LOTUS: 如果有一个随机变量 Y 是另一个随机变量 X 的函数：$Y = g(X)$，那么

$$\mathbb{E}_{p_Y(y)}[Y] = \mathbb{E}_{p_X(x)}[g(X)]$$

如果将 T 看作 g，$Y = z$

由此可以改写原式：

$$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\,x \mid z\,)] = \mathbb{E}_{\pi(\epsilon)}[\log p_\theta(x \mid T_\phi(x,\epsilon))]$$

这时进行 MC 时，则会变成

$$\mathbb{E}_{\pi(\epsilon)}[\log p_\theta(x \mid T_\phi(x,\epsilon))] \quad \epsilon_i \sim \mathcal{N}(\epsilon; 0, I)$$

$$\nabla_\phi \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\, x \mid z \,)] = \mathbb{E}_{\pi(\epsilon)}\left[\nabla_\phi \log p_\theta \left(\, x \, \middle| \, T_\phi(x, \epsilon) \, \right)\right]$$
$$= \mathbb{E}_{\pi(\epsilon)}\left[\nabla_\phi z \nabla_z \log p_\theta(\, x \mid z \,)\right]|_{z=T_\phi(x,\epsilon)}$$
$$\approx \nabla_\phi z \nabla_z \log p_\theta(\, x \mid z \,)\,|_{z=T_\phi(x,\epsilon)} \qquad \epsilon \sim \pi(\epsilon)$$

至此

$$\mathcal{L}(\phi, \theta) = \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\, x \mid z \,)]}_{\text{Reconstruction Error}} - \underbrace{KL[q_\phi(z \mid x) \parallel p(z)]}_{\text{Regularizer}}$$
$$\approx \frac{1}{M}\sum_{m=1}^{M}\left\{\log p_\theta \left(x_m \, \middle| \, T_\phi(x, \epsilon) \right) - KL[q_\phi(z_m \mid x_m) \parallel p(z_m)]\right\} \quad x_i \sim \mathcal{X}, \epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Summary

- Combining all the ingredients together:

$$\theta^*, \phi^* = argmax\ L(\phi, \theta)$$

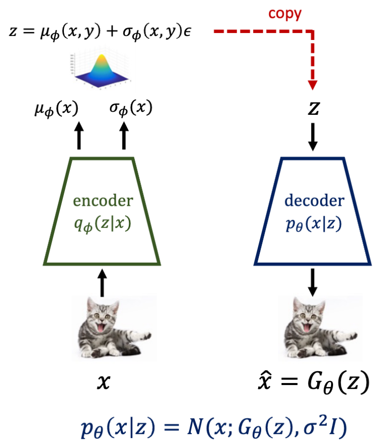
Reconstruction loss

$$L(\phi,\theta) \coloneqq E_{p_{data}(x)}\{ \boxed{-E_{N(\epsilon;0,I)}\left[\frac{1}{2\sigma^2} \parallel \underline{G_\theta\left(T_\phi(x,\epsilon)\right)} - x \parallel_2^2\right]}$$

stochastic auto-encoder

$$\underline{-KL[q_\phi(z|x) \parallel p(z)]}\ \}$$

KL regularizer
to make q closer to the prior and prevent $\sigma_\phi(x) \to 0$



Practical implementation for solving (pseudo code):

$$\max_{\theta,\phi} E_{p_{data}(x)}[E_{q_\phi(z|x)}[\underbrace{\log p_\theta(x|z)}_{= -\frac{1}{2\sigma^2}\|x - G_\theta(z)\|_2^2 + C}] - KL[q_\phi(z|x) \parallel p(z)]]$$

- Initialise θ, ϕ, learning rates γ, choose total iteration T for SGD
- For $t = 1, \dots, T$
  - $x_1, \dots, x_M \sim p_{data}(x)$

  \# encoder: performing (approximate) posterior inference
  - Compute $\mu_\phi(x_m), \sigma_\phi(x_m)$ for $m = 1, \dots, M$
  - $z_m = \mu_\phi(x_m) + \sigma_\phi(x_m) \odot \epsilon_m,\ \epsilon_m \sim N(0, I)$ \# reparam. trick

  \# Decoder: reconstructing data
  - $\hat{x}_m = G_\theta(z_m)$ for $m = 1, \dots, M$

  \# update neural network parameters
  - $L = \frac{1}{M}\sum_{m=1}^{M}[-\frac{1}{2\sigma^2}\|x_m - \hat{x}_m\|_2^2 - \underline{KL[q_\phi(z_m|x_m) \parallel p(z_m)]}]$
  - $(\theta, \phi) \leftarrow (\theta, \phi) + \gamma\, \nabla_{(\theta,\phi)} L$

can use the analytic KL form
or estimated by Monte Carlo

A practical trick: KL annealing

Section 3: Conditional VAE

假设额外的信息为 y，则 Latent Variable Model 被定义为

$$p_\theta(x \mid y) = \int_z p_\theta(\, x \mid z, y \,) p(z) dx$$

通常 $p(z) = \mathcal{N}(z; 0, I)$，如果 x 是连续，则有

$$p_\theta(\, x \mid z, y \,) = \mathcal{N}(x; G_\theta(z, y), \sigma^2 I)$$

类似的，我们优化 ELBO

$$\phi^*, \theta^* = \underset{\phi,\theta}{\operatorname{argmax}}\, \mathcal{L}(\phi, \theta)$$

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{p_{data}(x,y)} \Big\{ \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(\, x \mid z, y \,)] - KL[q_\phi(z \mid x, y) \parallel p(z)] \Big\}$$

尽管 q 的选择是自由的，但使用$q_\phi(z|x, y)$ 并用灵活的神经网络对其进行参数化将得到最佳的后验近似。

$$p_\theta(\, z \mid x, y \,) = \frac{p_\theta(x \mid z, y) p(z)}{p_\theta(x \mid y)}$$

关于 q 最大化 ELMO 等价于最小化 KL 散度 $KL[q_\phi(\, z \mid x, y \,) \parallel p_\theta(z \mid x, y)]$：

$$\begin{aligned}
&\log p_\theta(\, x \mid y \,) - \overbrace{\Big(\mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(\, x \mid z, y \,)] - KL[q_\phi(z \mid x, y) \parallel p(z)] \Big)}^{\text{ELBO}} \\
&= \log p_\theta(\, x \mid y \,) - \ \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(\, x \mid z, y \,)] + \mathbb{E}_{q_\phi(z|x,y)} \left[\log \frac{q_\phi(z \mid x, y)}{p(z)} \right] \\
&= \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(\, x \mid y \,)] + \mathbb{E}_{q_\phi(z|x,y)} \left[\log \frac{q_\phi(z \mid x, y)}{p(z) p_\theta(\, x \mid z, y \,)} \right] \\
&= \mathbb{E}_{q_\phi(z|x,y)} \left[\log \frac{p_\theta(\, x \mid y \,) q_\phi(z \mid x, y)}{p(z) p_\theta(\, x \mid z, y \,)} \right] \\
&= \mathbb{E}_{q_\phi(z|x,y)} \left[\log \frac{p_\theta(\, x \mid y \,) q_\phi(z \mid x, y) p(z)}{p(z) p_\theta(\, z \mid x, y \,) p_\theta(\, x \mid y \,)} \right] \\
&= \mathbb{E}_{q_\phi(z|x,y)} \left[\log \frac{q_\phi(z \mid x, y)}{p_\theta(\, z \mid x, y \,)} \right] \\
&= KL[q_\phi(z \mid x, y) \parallel p_\theta(\, z \mid x, y \,)]
\end{aligned}$$

如果我们用$q_\phi(z \mid x)$ 替代$q_\phi(z \mid x, y)$，那么除非学习到的生成器退化（degenerate）：$G_\theta(z, y) = G_\theta(z)$，否则最优解不会得到精确的后验近似。在这种情况下，y 信息被忽略（即$p_\theta(x|z, y) = p_\theta(x|z)$），模型不再是条件生成模型。