VARIATIONAL AUTOENCODER

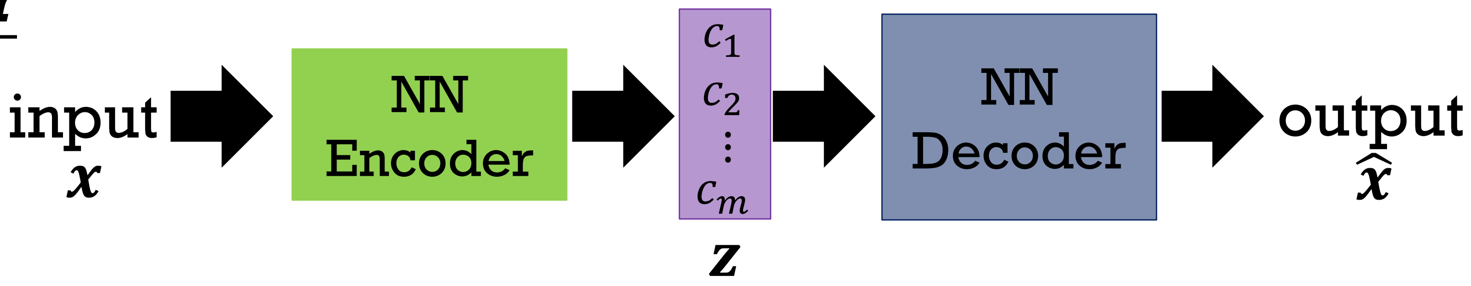
Pei-Yuan Wu

National Taiwan University

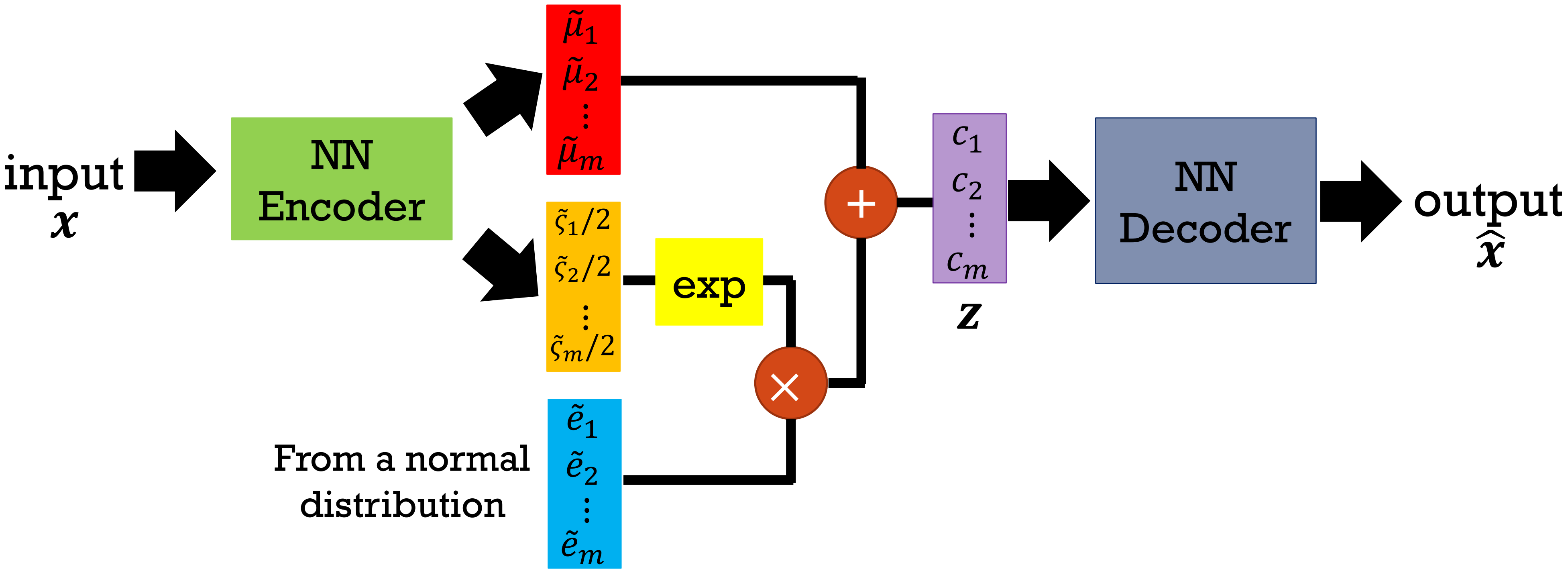
1

AUTO ENCODER V.S. VAE

Auto-encoder

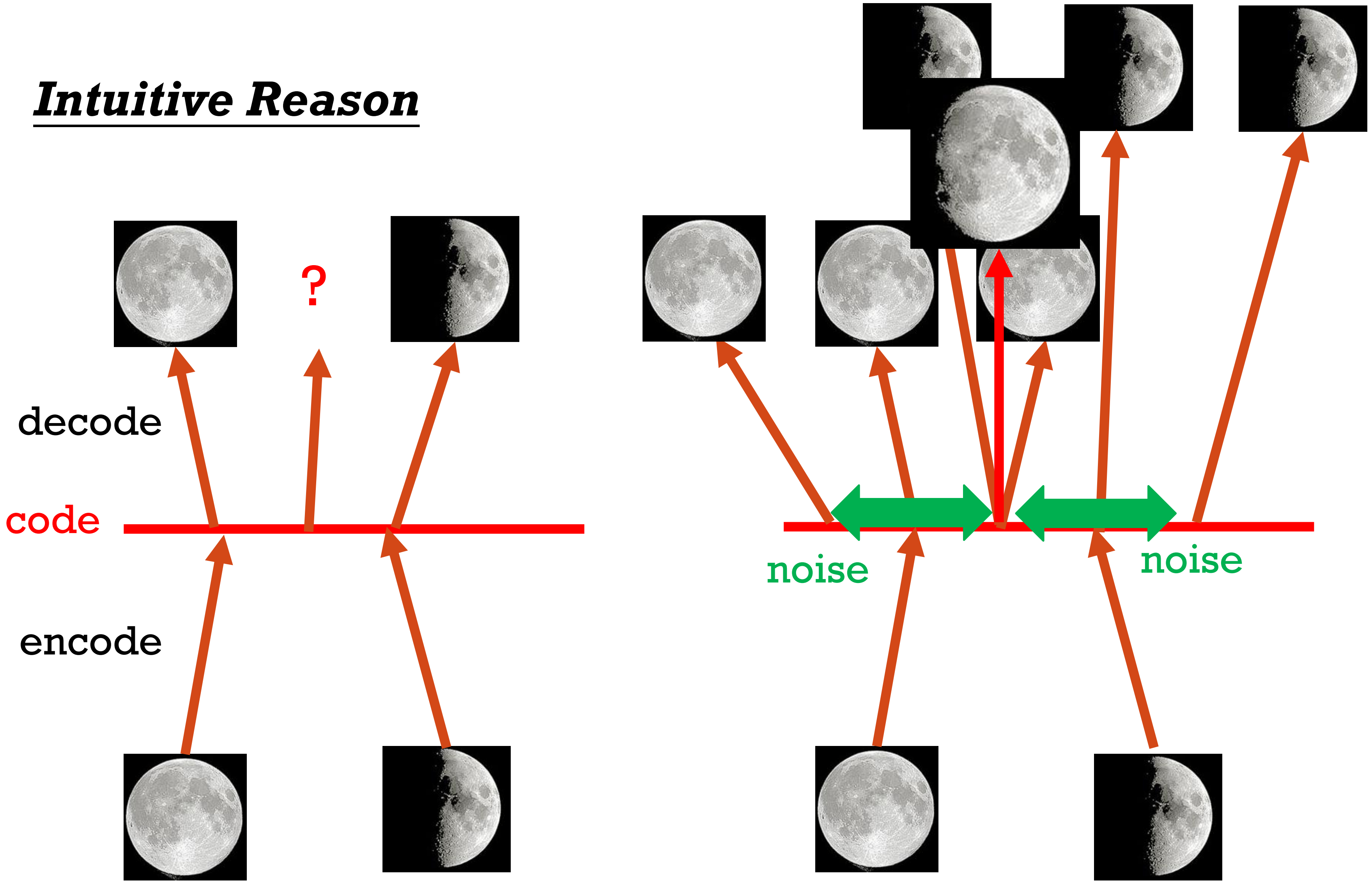


VAE



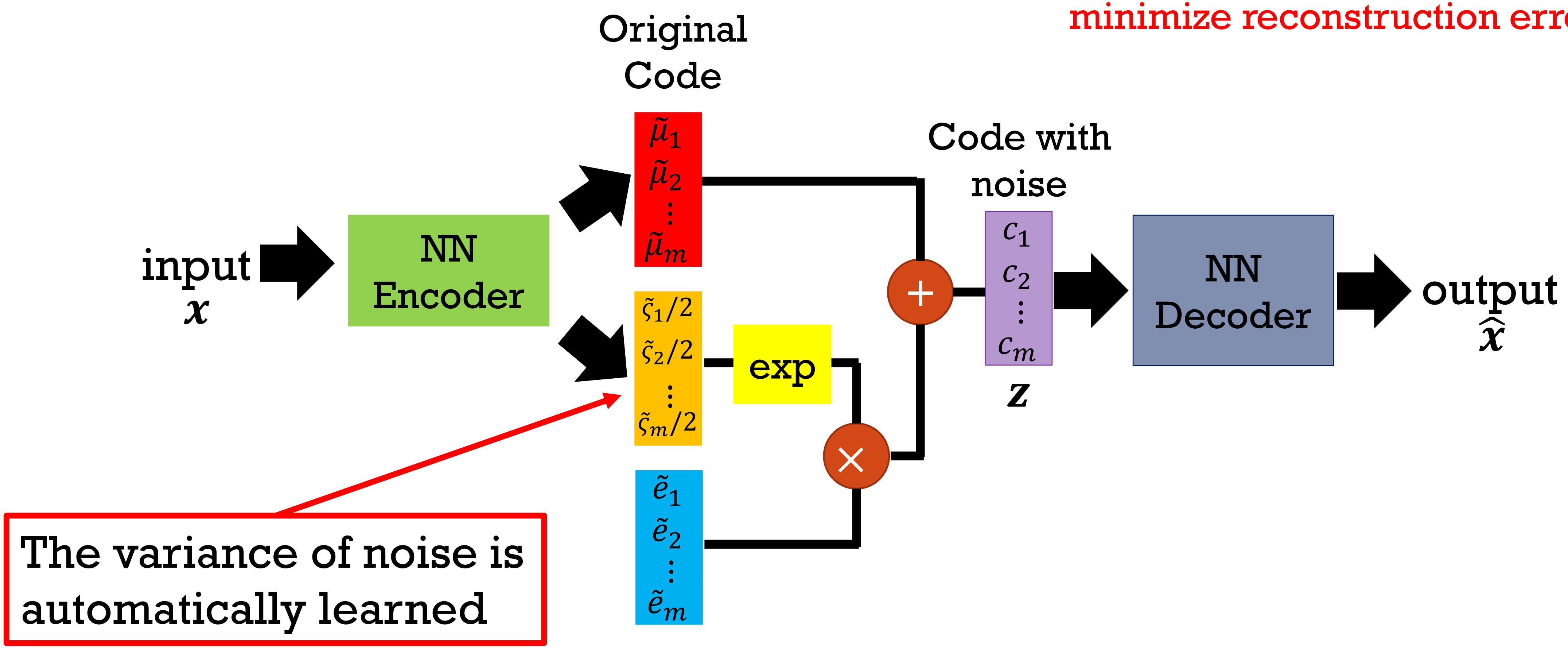
WHY VAE?

Intuitive Reason



VAE LOSS FUNCTION

What will happen if we only minimize reconstruction error?



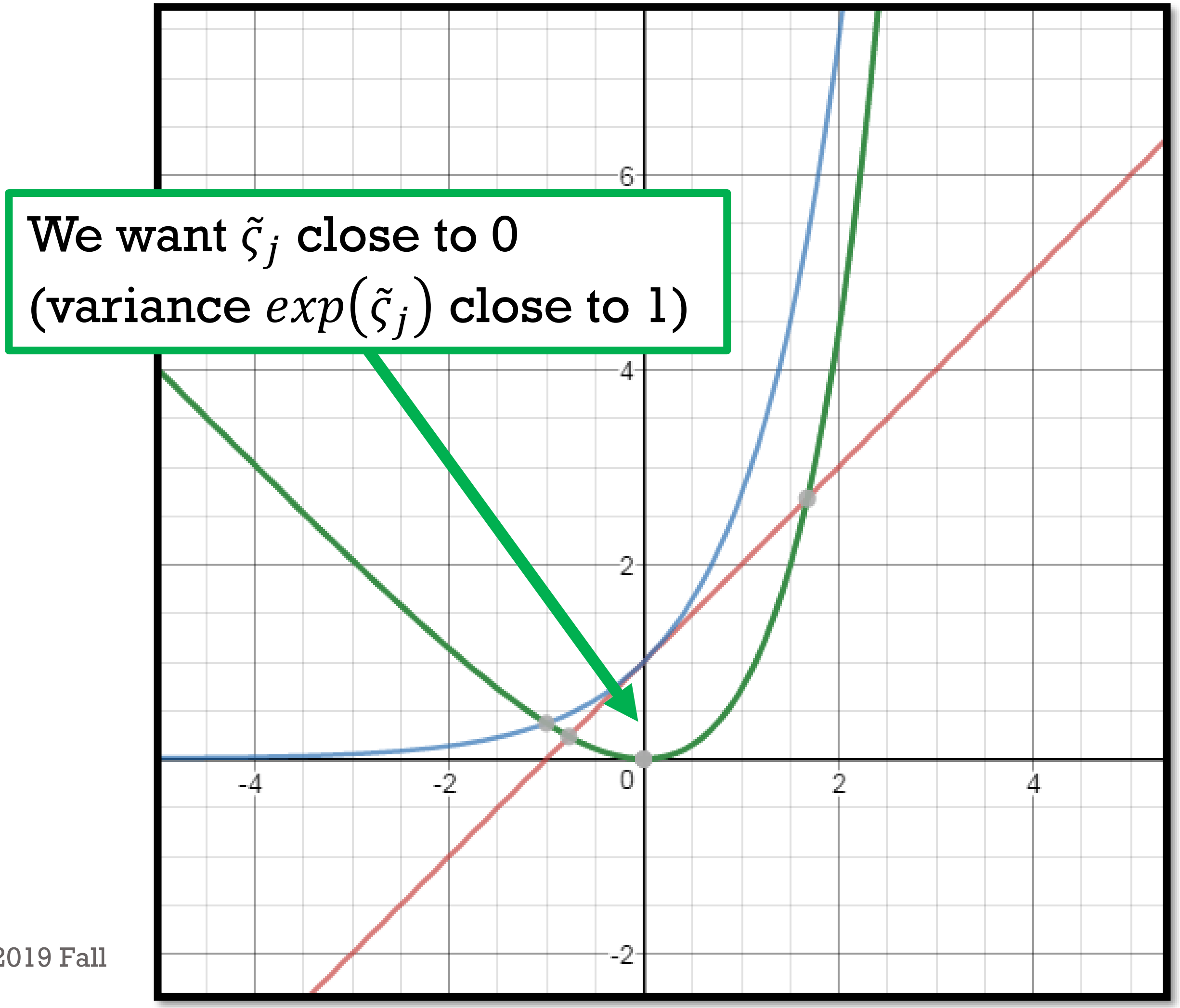
Given data set $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, minimize

$$L_{VAE} = \sum_{i=1}^{N} \left(\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2 + \sum_{j=1}^{m} \left(\left\|\tilde{\mu}_j(\boldsymbol{x}_i)\right\|^2 + exp\left(\tilde{\varsigma}_j(\boldsymbol{x}_i)\right) - \left(1 + \tilde{\varsigma}_j(\boldsymbol{x}_i)\right)\right)\right)$$

Reconstruction loss　　Regularization loss

WHY VAE REGULARIZATION?

Intuitive Reason: *Want the code to have zero mean and unit variance*



Regularization loss

$$\left\|\tilde{\mu}_j(\boldsymbol{x}_i)\right\|^2 + exp\left(\tilde{\varsigma}_j(\boldsymbol{x}_i)\right) - \left(1 + \tilde{\varsigma}_j(\boldsymbol{x}_i)\right)$$

VAE GENERATIVE MODEL AND MLE

- We have data set $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$, and assume each data point $\boldsymbol{x}_i$ is generated according to generative model

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\theta(\boldsymbol{z})d\boldsymbol{z}$$

- Goal: Maximum log-likelihood parameter estimation:

$$\theta^* = \max_{\theta\in\Theta} p_\theta(\mathcal{X}) = \max_{\theta\in\Theta} \log p_\theta(\mathcal{X})$$

  - If each data point $\boldsymbol{x}_i$ is generated independently, then

$$\log p_\theta(\mathcal{X}) = \log\left(\prod_{i=1}^{N} p_\theta(\boldsymbol{x}_i)\right) = \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i)$$

$$p_\theta(\boldsymbol{x}_i) = \int p_\theta(\boldsymbol{x}_i|\boldsymbol{z}_i)p_\theta(\boldsymbol{z}_i)d\boldsymbol{z}_i$$

- Introduce latent variables $\mathcal{Z} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_N\}$, where $\boldsymbol{z}_i$ indicates the latent code of $\boldsymbol{x}_i$.

Latent variable (not directly observable) $\boldsymbol{z} \in \mathbb{R}^m$

$p_\theta(\boldsymbol{x}|\boldsymbol{z})$

Data point (directly observable) $\boldsymbol{x} \in \mathbb{R}^M$

LOG-LIKELIHOOD LOWER BOUND

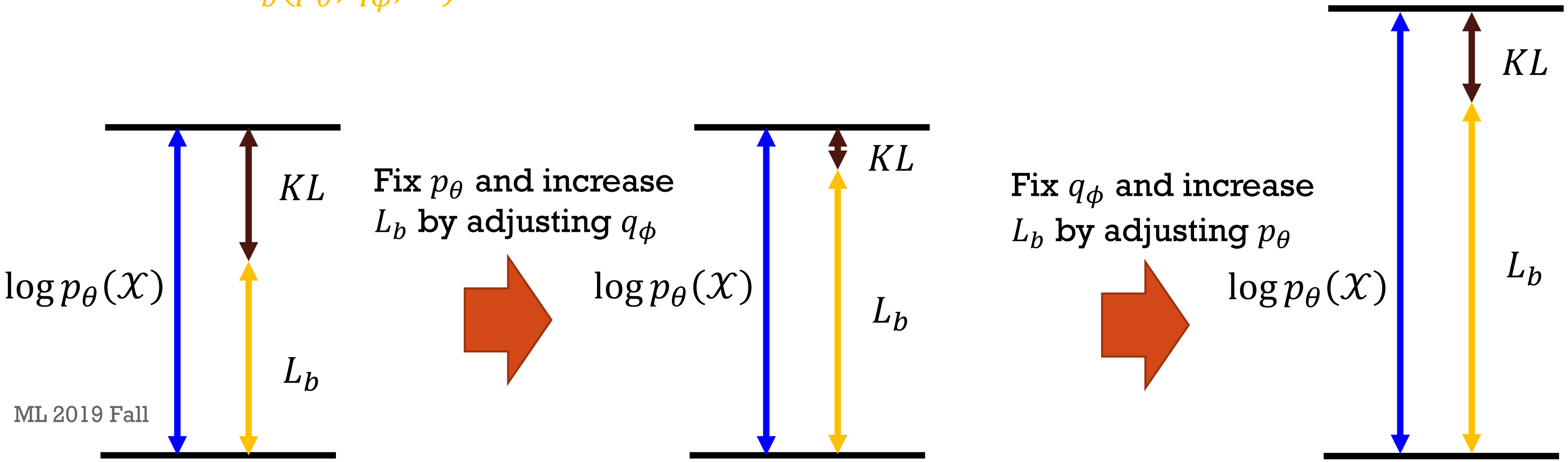
$$\log p_\theta(\mathcal{X}) = \int q_\phi(\mathcal{Z}|\mathcal{X}) \log p_\theta(\mathcal{X})\, d\mathcal{Z}$$

$$= \int q_\phi(\mathcal{Z}|\mathcal{X}) \left(\log \frac{p_\theta(\mathcal{Z},\mathcal{X})}{q_\phi(\mathcal{Z}|\mathcal{X})} - \log \frac{p_\theta(\mathcal{Z}|\mathcal{X})}{q_\phi(\mathcal{Z}|\mathcal{X})} \right) d\mathcal{Z}$$

$$= \int q_\phi(\mathcal{Z}|\mathcal{X}) \log \frac{p_\theta(\mathcal{Z},\mathcal{X})}{q_\phi(\mathcal{Z}|\mathcal{X})} d\mathcal{Z} + KL\left(q_\phi(\cdot\,|\mathcal{X}), p_\theta(\cdot\,|\mathcal{X})\right)$$

Tractable lower bound $L_b(p_\theta, q_\phi, \mathcal{X})$

Intractable but always ≥ 0



VAE V.S. EM ALGORITHM

- Randomly initialize parameters $\theta^{(1)}$.
- Iterate through step t=1,2,…
  - Expectation Step (E-step): Compute
    $$Q(\theta|\theta^{(t)}) = \sum_{\mathcal{Z}} p(\mathcal{Z}|\mathcal{X};\theta^{(t)}) \log p(\mathcal{X},\mathcal{Z};\theta)$$
  - Maximization Step (M-step): Choose
    $$\theta^{(t+1)} = \underset{\theta\in\Theta}{argmax}\, Q(\theta|\theta^{(t)})$$

Fix q_ϕ and adjust p_θ to maximize

$$L_b(p_\theta, q_\phi, \mathcal{X}) = \mathbb{E}_{\mathcal{Z}\sim q_\phi(\cdot|\mathcal{X})}\left[\log \frac{p_\theta(\mathcal{X},\mathcal{Z})}{q_\phi(\mathcal{Z}|\mathcal{X})}\right]$$

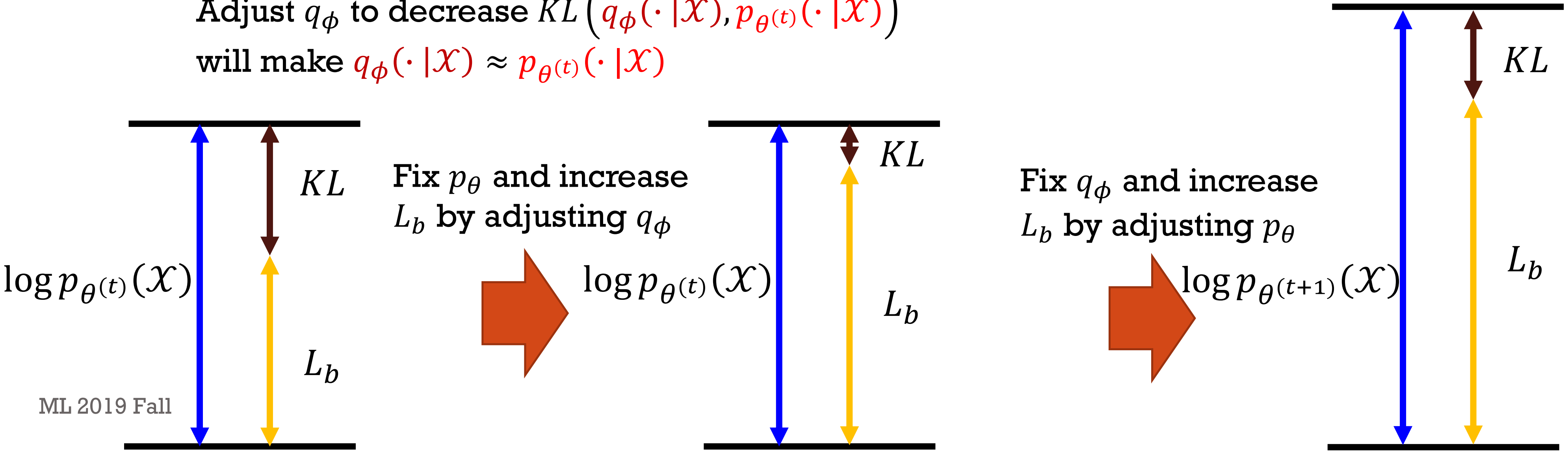
is equivalent to maximizing

$$\mathbb{E}_{\mathcal{Z}\sim q_\phi(\cdot|\mathcal{X})}[\log p_\theta(\mathcal{X},\mathcal{Z})] \approx Q(\theta|\theta^{(t)})$$

approx. $p_{\theta^{(t)}}(\cdot|\mathcal{X})$

EM Algorithm

Adjust q_ϕ to decrease $KL\left(q_\phi(\cdot|\mathcal{X}), p_{\theta^{(t)}}(\cdot|\mathcal{X})\right)$ will make $q_\phi(\cdot|\mathcal{X}) \approx p_{\theta^{(t)}}(\cdot|\mathcal{X})$



VAE WITH INDEPENDENT SAMPLES

- Goal: Adjust p_θ and q_ϕ to maximize

$$L_b(p_\theta, q_\phi, \mathcal{X}) = \mathbb{E}_{\mathcal{Z}\sim q_\phi(\cdot|\mathcal{X})}\left[\log \frac{p_\theta(\mathcal{X},\mathcal{Z})}{q_\phi(\mathcal{Z}|\mathcal{X})}\right]$$

$$= \mathbb{E}_{\mathcal{Z}\sim q_\phi(\cdot|\mathcal{X})}\left[\log \frac{p_\theta(\mathcal{X}|\mathcal{Z})}{q_\phi(\mathcal{Z}|\mathcal{X})}\right] + \mathbb{E}_{\mathcal{Z}\sim q_\phi(\cdot|\mathcal{X})}[\log p_\theta(\mathcal{Z})]$$

- Assume $(\boldsymbol{x}_1, \boldsymbol{z}_1), \dots, (\boldsymbol{x}_N, \boldsymbol{z}_N)$ are independent w.r.t. p_θ and q_ϕ, then

$$q_\phi(\mathcal{Z}|\mathcal{X}) = \prod_{i=1}^N q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i), p_\theta(\mathcal{X}|\mathcal{Z}) = \prod_{i=1}^N p_\theta(\boldsymbol{x}_i|\boldsymbol{z}_i), p_\theta(\mathcal{Z}) = \prod_{i=1}^N p_\theta(\boldsymbol{z}_i)$$

$$\mathbb{E}_{\mathcal{Z}\sim q_\phi(\cdot|\mathcal{X})}\left[\log \frac{p_\theta(\mathcal{X}|\mathcal{Z})}{q_\phi(\mathcal{Z}|\mathcal{X})}\right] = \sum_{i=1}^N \mathbb{E}_{\mathcal{Z}\sim q_\phi(\cdot|\mathcal{X})}\left[\log \frac{p_\theta(\boldsymbol{x}_i|\boldsymbol{z}_i)}{q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i)}\right] = \sum_{i=1}^N \mathbb{E}_{\boldsymbol{z}_i\sim q_\phi(\cdot|\boldsymbol{x}_i)}\left[\log \frac{p_\theta(\boldsymbol{x}_i|\boldsymbol{z}_i)}{q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i)}\right]$$

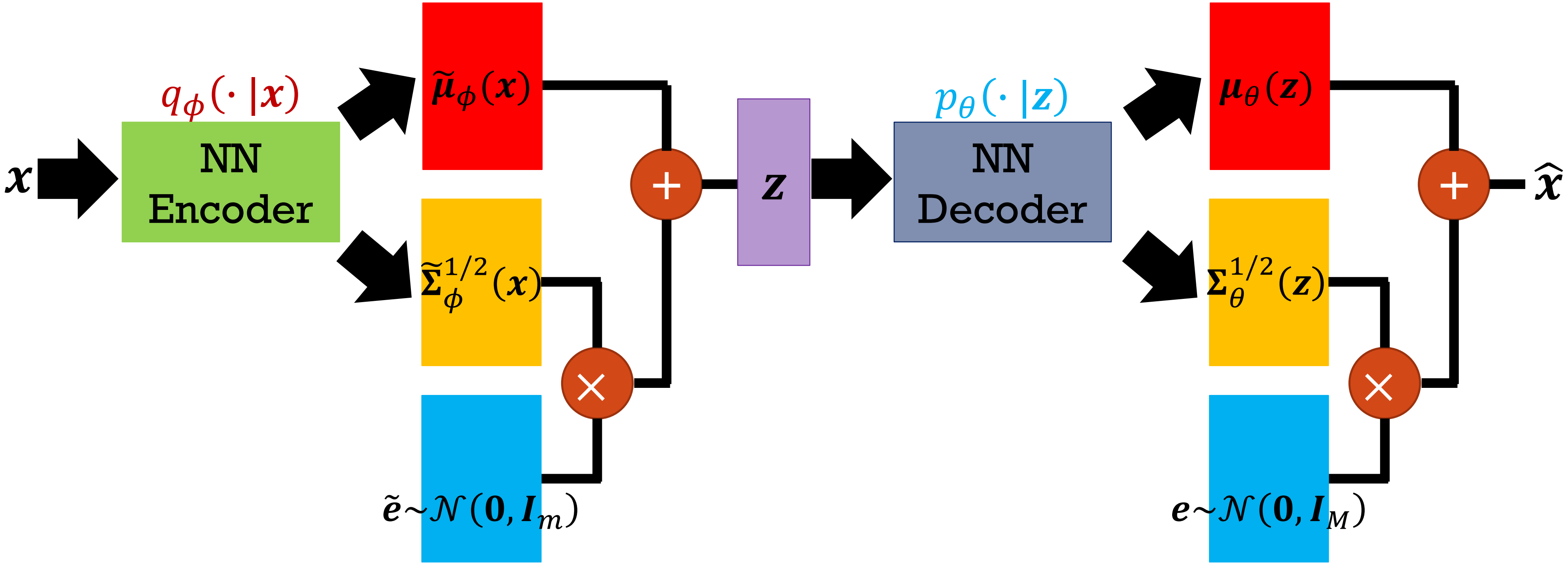
$$\mathbb{E}_{\mathcal{Z}\sim q_\phi(\cdot|\mathcal{X})}[\log p_\theta(\mathcal{Z})] = \sum_{i=1}^N \mathbb{E}_{\mathcal{Z}\sim q_\phi(\cdot|\mathcal{X})}[\log p_\theta(\boldsymbol{z}_i)] = \sum_{i=1}^N \mathbb{E}_{\boldsymbol{z}_i\sim q_\phi(\cdot|\boldsymbol{x}_i)}[\log p_\theta(\boldsymbol{z}_i)]$$

Hence

$$L_b(p_\theta, q_\phi, \mathcal{X}) = \sum_{i=1}^N \mathbb{E}_{\boldsymbol{z}_i\sim q_\phi(\cdot|\boldsymbol{x}_i)}\left[\log \frac{p_\theta(\boldsymbol{x}_i|\boldsymbol{z}_i)}{q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i)} + \log p_\theta(\boldsymbol{z}_i)\right]$$

VAE AND GAUSSIAN DISTRIBUTION

- Assume latent code has Gaussian prior, and both encoder/decoder described by Gaussian distributions: $p_\theta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$, $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z}))$, $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x})\right)$.



CLOSE FORM OF L_b

- Assume latent code has Gaussian prior, and both encoder/decoder described by Gaussian distributions: $p_\theta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$, $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_\theta(\boldsymbol{z}), \boldsymbol{\Sigma}_\theta(\boldsymbol{z}))$, $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x})\right)$. Then

$$\mathbb{E}_{\boldsymbol{z}_i \sim q_\phi(\cdot|\boldsymbol{x}_i)}\left[\log \frac{p_\theta(\boldsymbol{x}_i|\boldsymbol{z}_i)}{q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i)} + \log p_\theta(\boldsymbol{z}_i)\right] = \mathbb{E}_{\boldsymbol{z}_i \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i))}\left[\log \frac{\mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_\theta(\boldsymbol{z}_i), \boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i))}{\mathcal{N}\left(\boldsymbol{z}_i; \widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i)\right)} + \log \mathcal{N}(\boldsymbol{z}_i; \boldsymbol{0}, \boldsymbol{I})\right]$$

$$= \mathbb{E}_{\boldsymbol{z}_i \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i))}[\log \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_\theta(\boldsymbol{z}_i), \boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i))] + \log \frac{\mathcal{N}\left(\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i); \boldsymbol{0}, \boldsymbol{I}\right)}{\mathcal{N}\left(\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i); \widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i)\right)} - \frac{Tr\left(\widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i) - \boldsymbol{I}\right)}{2}$$

$$= \mathbb{E}_{\boldsymbol{z}_i \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i)\right)}\left[\log\left(\frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i)|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_\theta(\boldsymbol{z}_i)\right)^T \boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i)^{-1}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_\theta(\boldsymbol{z}_i)\right)\right)\right)\right]$$

$$+ \log\left(\frac{\frac{1}{\sqrt{(2\pi)^m}} \exp\left(-\frac{1}{2}\left\|\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i)\right\|^2\right)}{1/\sqrt{(2\pi)^m |\widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i)|}}\right) - \frac{Tr\left(\widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i) - \boldsymbol{I}\right)}{2}$$

$$= \mathbb{E}_{\boldsymbol{z}_i \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i)\right)}\left[\log \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i)|}} - \frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_\theta(\boldsymbol{z}_i)\right)^T \boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i)^{-1}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_\theta(\boldsymbol{z}_i)\right)\right] + \frac{1}{2}\log|\boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i)|$$

$$- \frac{1}{2}\left\|\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i)\right\|^2 - \frac{Tr\left(\widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i) - \boldsymbol{I}\right)}{2}$$

Lemma: Let $\boldsymbol{\xi} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ be a Gaussian-distributed r.v. in $\mathbb{R}^m$, then

$$\mathbb{E}_{\xi\sim\mathcal{N}(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})}[\log\mathcal{N}(\boldsymbol{\xi};\boldsymbol{\mu},\boldsymbol{\Sigma})] = \log\frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} - \frac{1}{2}\left((\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}) + Tr\left(\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\right)\right)$$

$$= \log\mathcal{N}(\tilde{\boldsymbol{\mu}};\boldsymbol{\mu},\boldsymbol{\Sigma}) - Tr\left(\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\right)/2$$

In particular, if $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}$, $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$, then

$$\mathbb{E}_{\xi\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\log\mathcal{N}(\boldsymbol{\xi};\boldsymbol{\mu},\boldsymbol{\Sigma})] = -\frac{m}{2}\log\left(2\pi e|\boldsymbol{\Sigma}|^{1/m}\right)$$

Proof:

$$\mathbb{E}_{\xi\sim\mathcal{N}(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})}[\log\mathcal{N}(\boldsymbol{\xi};\boldsymbol{\mu},\boldsymbol{\Sigma})] = \mathbb{E}_{\xi\sim\mathcal{N}(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})}\left[\log\frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} - \frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu})\right]$$

$$= \log\frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} - \frac{1}{2}\mathbb{E}_{\xi\sim\mathcal{N}(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})}\left[Tr\left((\boldsymbol{\xi}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu})\right)\right]$$

$$= \log\frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} - \frac{1}{2}Tr\left(\mathbb{E}_{\xi\sim\mathcal{N}(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})}[(\boldsymbol{\xi}-\boldsymbol{\mu})(\boldsymbol{\xi}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}]\right)$$

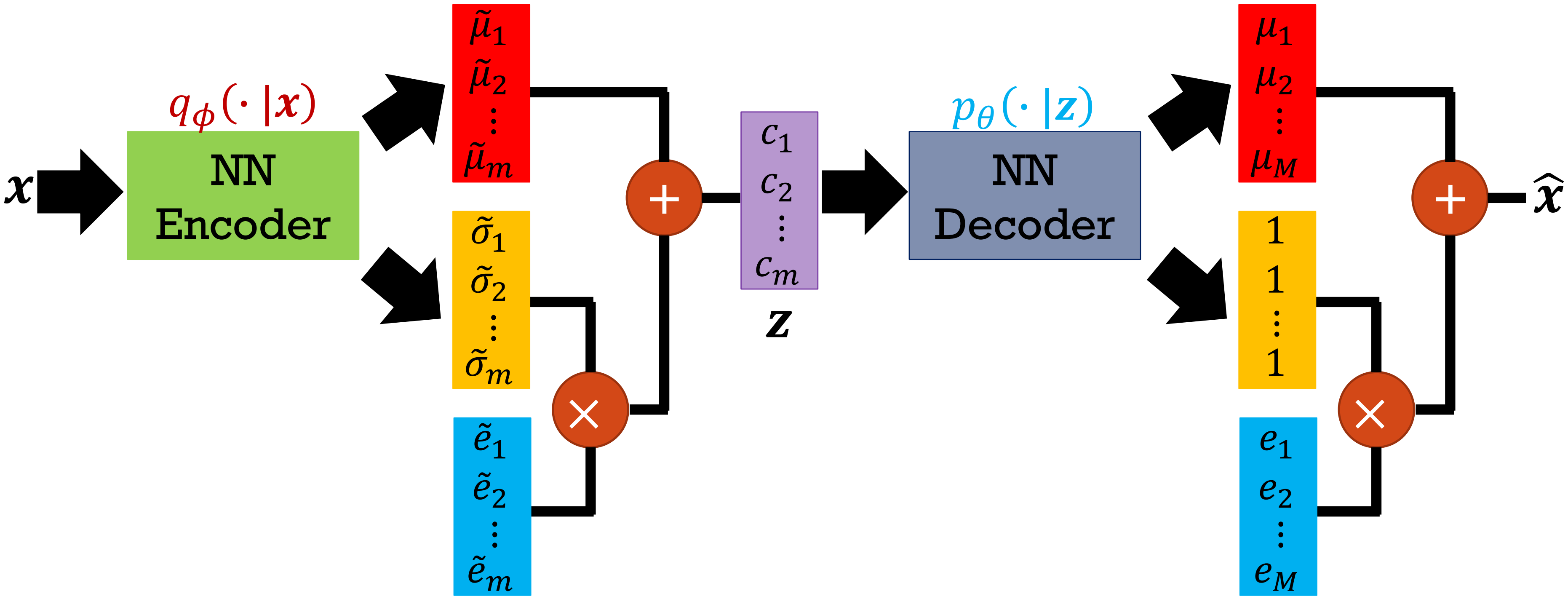
$$= \log\frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} - \frac{1}{2}Tr\left(\mathbb{E}_{\xi\sim\mathcal{N}(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})}[(\boldsymbol{\xi}-\boldsymbol{\mu})(\boldsymbol{\xi}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}]\right)$$

$$= \log\frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} - \frac{1}{2}Tr\left(\mathbb{E}_{\xi'\sim\mathcal{N}(\mathbf{0},\tilde{\boldsymbol{\Sigma}})}[(\boldsymbol{\xi}'+\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu})(\boldsymbol{\xi}'+\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}]\right)$$

$$= \log\frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} - \frac{1}{2}Tr\left(\left(\tilde{\boldsymbol{\Sigma}} + (\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu})(\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu})^T\right)\boldsymbol{\Sigma}^{-1}\right) = \log\frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} - \frac{1}{2}\left((\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}) + Tr\left(\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\right)\right)$$

SIMPLER CLOSE FORM OF L_b

- For simplicity, assume $\boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i) = \boldsymbol{I}, \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}) = diag(\tilde{\sigma}^2_{\phi,1}(\boldsymbol{x}), \dots, \tilde{\sigma}^2_{\phi,m}(\boldsymbol{x}))$.



VAE LOSS FUNCTION

- For simplicity, assume $\boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i) = \boldsymbol{I}, \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}) = diag(\tilde{\sigma}^2_{\phi,1}(\boldsymbol{x}), \dots, \tilde{\sigma}^2_{\phi,m}(\boldsymbol{x}))$, then

$$\mathbb{E}_{\boldsymbol{z}_i \sim q_\phi(\cdot|\boldsymbol{x}_i)}\left[\log \frac{p_\theta(\boldsymbol{x}_i|\boldsymbol{z}_i)}{q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i)} + \log p_\theta(\boldsymbol{z}_i)\right]$$

$$= \mathbb{E}_{\boldsymbol{z}_i \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i))}\left[\log \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i)|}} - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_\theta(\boldsymbol{z}_i))^T \boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i)^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_\theta(\boldsymbol{z}_i))\right] + \frac{1}{2}\log|\boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i)| - \frac{1}{2}\left\|\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i)\right\|^2$$

$$- \frac{1}{2}Tr\left(\widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i) - \boldsymbol{I}\right) = \log \frac{1}{\sqrt{(2\pi)^M}} - \frac{1}{2}\left(\mathbb{E}_{\boldsymbol{z}_i \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i)\right)}[\|\boldsymbol{x}_i - \boldsymbol{\mu}_\theta(\boldsymbol{z}_i)\|^2] + \left\|\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i)\right\|^2 + \sum_{j=1}^m \left(\tilde{\sigma}^2_{\phi,j}(\boldsymbol{x}_i) - \log \tilde{\sigma}^2_{\phi,j}(\boldsymbol{x}_i) - 1\right)\right)$$

Hence maximizing

$$L_b\left(p_\theta, q_\phi, \mathcal{X}\right) = N \log \frac{1}{\sqrt{(2\pi)^M}} - \frac{1}{2}\sum_{i=1}^N \left(\mathbb{E}_{\boldsymbol{z}_i \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i)\right)}[\|\boldsymbol{x}_i - \boldsymbol{\mu}_\theta(\boldsymbol{z}_i)\|^2] + \left\|\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i)\right\|^2 + \sum_{j=1}^m \left(\tilde{\sigma}^2_{\phi,j}(\boldsymbol{x}_i) - \log \tilde{\sigma}^2_{\phi,j}(\boldsymbol{x}_i) - 1\right)\right)$$

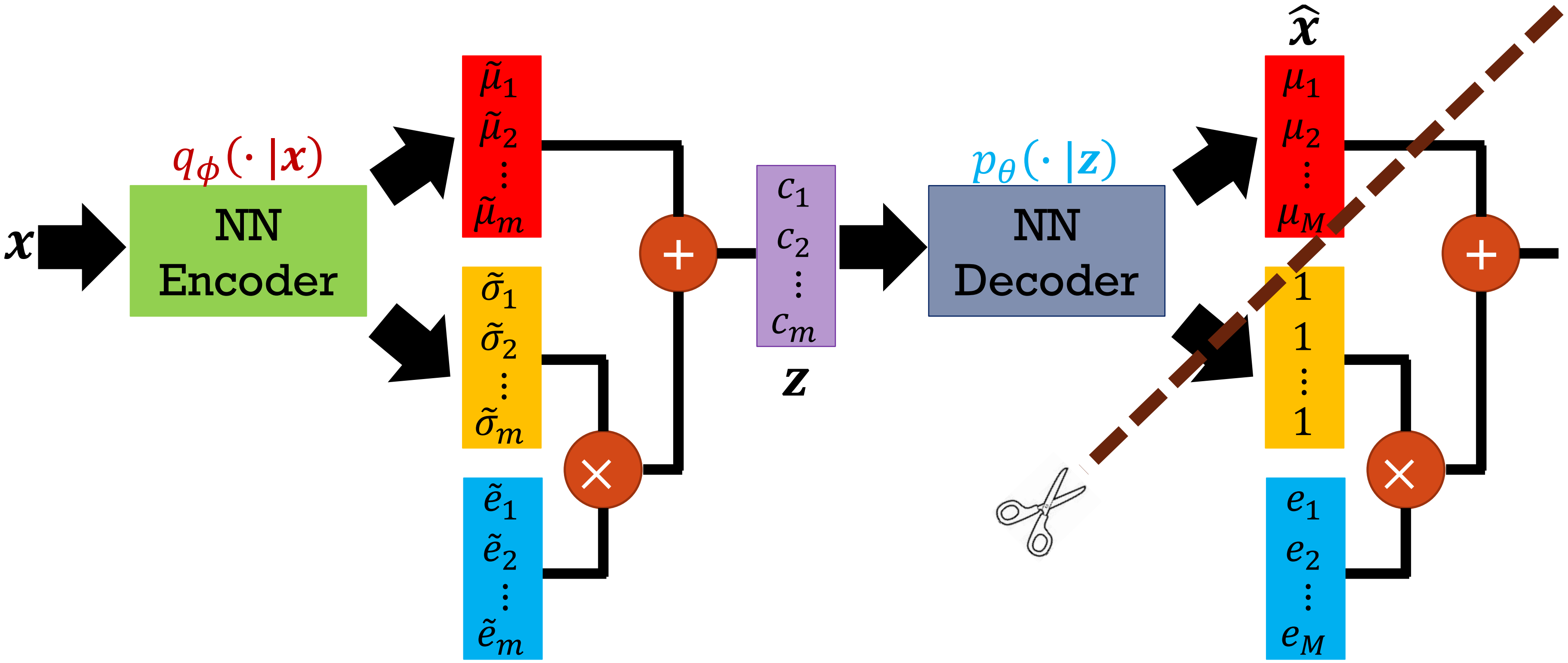
Is equivalent to minimizing

$$\sum_{i=1}^N \left(\mathbb{E}_{\boldsymbol{z}_i \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i), \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}_i)\right)}[\|\boldsymbol{x}_i - \boldsymbol{\mu}_\theta(\boldsymbol{z}_i)\|^2] + \left\|\widetilde{\boldsymbol{\mu}}_\phi(\boldsymbol{x}_i)\right\|^2 + \sum_{j=1}^m \left(\tilde{\sigma}^2_{\phi,j}(\boldsymbol{x}_i) - \left(1 + \log \tilde{\sigma}^2_{\phi,j}(\boldsymbol{x}_i)\right)\right)\right)$$

Approx. expectation by sample mean

VAE LOSS FUNCTION

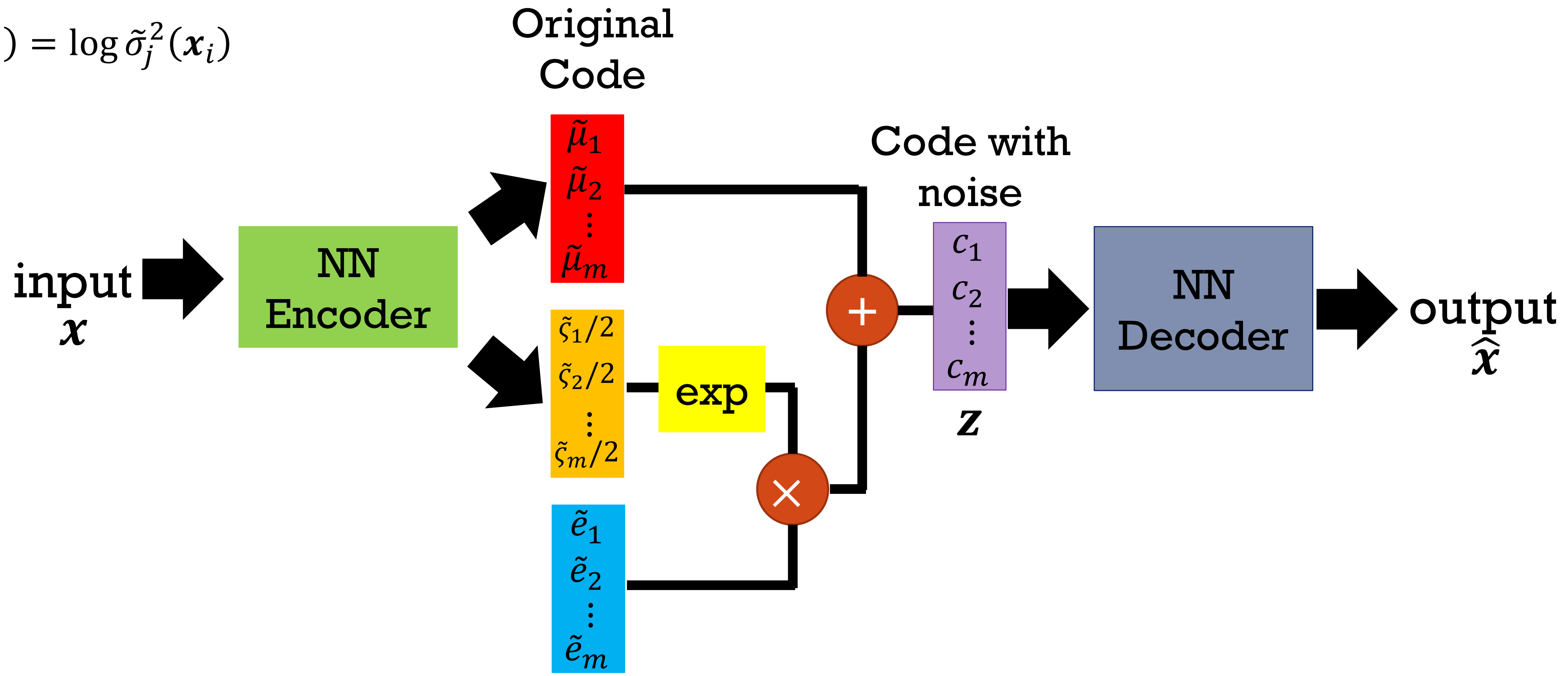
- For simplicity, assume $\boldsymbol{\Sigma}_\theta(\boldsymbol{z}_i) = \boldsymbol{I}, \widetilde{\boldsymbol{\Sigma}}_\phi(\boldsymbol{x}) = diag(\tilde{\sigma}^2_{\phi,1}(\boldsymbol{x}), \dots, \tilde{\sigma}^2_{\phi,m}(\boldsymbol{x}))$.



$$L_{VAE} = \sum_{i=1}^{N} \left(\|\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i\|^2 + \sum_{j=1}^{m} \left(\left\|\tilde{\mu}_j(\boldsymbol{x}_i)\right\|^2 + \tilde{\sigma}_j^2(\boldsymbol{x}_i) - \left(1 + \log \tilde{\sigma}_j^2(\boldsymbol{x}_i)\right) \right) \right)$$

VAE LOSS FUNCTION

Take $\tilde{\varsigma}_j(\boldsymbol{x}_i) = \log \tilde{\sigma}_j^2(\boldsymbol{x}_i)$



Minimize

$$L_{VAE} = \sum_{i=1}^{N} \left(\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2 + \sum_{j=1}^{m} \left(\left\|\tilde{\mu}_j(\boldsymbol{x}_i)\right\|^2 + exp\left(\tilde{\varsigma}_j(\boldsymbol{x}_i)\right) - \left(1 + \tilde{\varsigma}_j(\boldsymbol{x}_i)\right) \right) \right)$$