Autoencoders

"Manifold hypothesis" – data lives on a low-dimensional manifold

Dimensionality reduction

$x \in \mathbb{R}^D \rightarrow z \in \mathbb{R}^d \qquad d < D \quad (d \ll D)$

$f(x) = z$

– f – linear $\quad X \in \mathbb{R}^{N \times D}$ $\quad$ N – num of training objects

$Z \in \mathbb{R}^{N \times d} \qquad Z = X\Theta \qquad \Theta \in \mathbb{R}^{D \times d}$

Reconstruction $\quad \hat{X} = Z\Phi \qquad \Phi \in \mathbb{R}^{d \times D}$

$$X \approx \hat{X} = Z\Phi = X\Theta\Phi$$

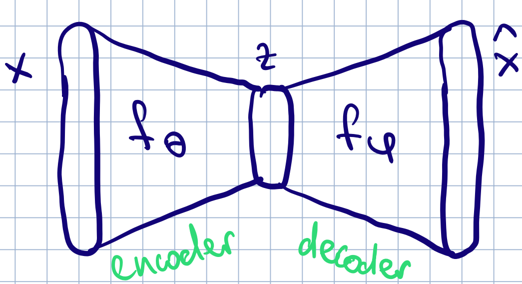
$$MSE = \|X - \hat{X}\|_F^2 = \|X - \underbrace{X \overset{D\times d}{\Theta} \overset{d \times D}{\Phi}}_{\text{rank } d}\|_F^2 \rightarrow \min_{\Theta, \Phi}$$

$$\|A\|_F = \sqrt{\sum_{ij} |A_{ij}|^2}$$

- Principal Component Analysis (PCA)

– f – non linear

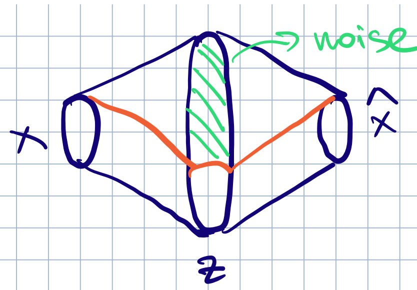
$\mathbb{X}$ – data distribution



$$\mathbb{E}_{x \sim \mathbb{X}} \|x - \hat{x}\|_2^2 =$$

$$= \mathbb{E}_{x \sim \mathbb{X}} \|x - f_\varphi(f_\theta(x))\|_2^2 \rightarrow \min_{\varphi, \theta}$$

$d > D$



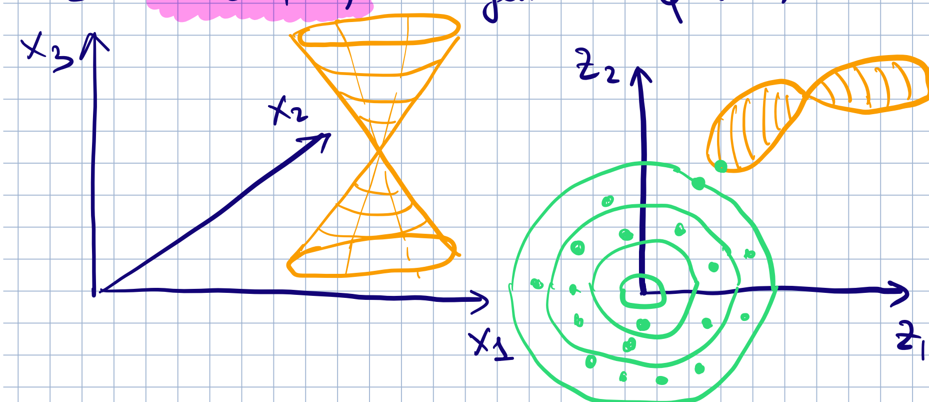
Denoising Autoencoder

$$\mathbb{E}_{x\sim \mathbb{X}} \| x - f_\varphi(f_\theta(\tilde{x})) \|_2^2 \to \min_{\varphi,\theta}$$

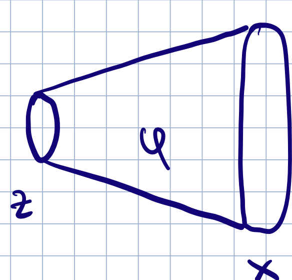
$$\underbrace{\tilde{x} = x + \varepsilon}, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \cdot I)$$

Sampling from Autoencoder

$$z \sim \mathcal{N}(0, I) \qquad x_{gen} = f_\varphi(z)$$



Variational Autoencoder (VAE)



$$p(x,z \mid \varphi) = p(x \mid z, \varphi)\, p(z)$$

$$p(x \mid z, \varphi) = \mathcal{N}(x \mid \underbrace{\mu_\varphi(z)}_{\in \mathbb{R}^D}, \underbrace{\Sigma_\varphi(z)}_{\in \mathbb{R}^{D\times D}})$$

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

x – observe $\quad z$ – latent

$$\log p(x|\varphi) = \log \int p(x,z|\varphi)\,dz \to \max_{\varphi}$$

Evidence lower Bound (ELBO)

$$\log p(x|\varphi) = \log p(x|\varphi)\cdot 1 = \left\{ q(z) - \text{arbitrary distribution over } z \right\}$$

$$= \log p(x|\varphi)\cdot \int q(z)\,dz = \int q(z) \log p(x|\varphi)\,dz =$$

$$= \left\{ p(x,z|\varphi) = p(z|x,\varphi)\cdot p(x|\varphi) \right\} =$$

$$= \int q(z) \log \frac{p(x,z|\varphi)}{p(z|x,\varphi)}\,dz =$$

$$= \int q(z) \log \frac{p(x,z|\varphi)\cdot q(z)}{p(z|x,\varphi)\cdot q(z)}\,dz =$$

$$= \underbrace{\int q(z) \log \frac{p(x,z|\varphi)}{q(z)}\,dz}_{} + \underbrace{\int q(z) \log \frac{q(z)}{p(z|x,\varphi)}\,dz}_{= KL(q\,\|\,p(z|x,\varphi))} =$$

$$= \underbrace{\mathcal{L}(q,\varphi)}_{\text{ELBO}} + KL(q\,\|\,p(z|x,\varphi))$$

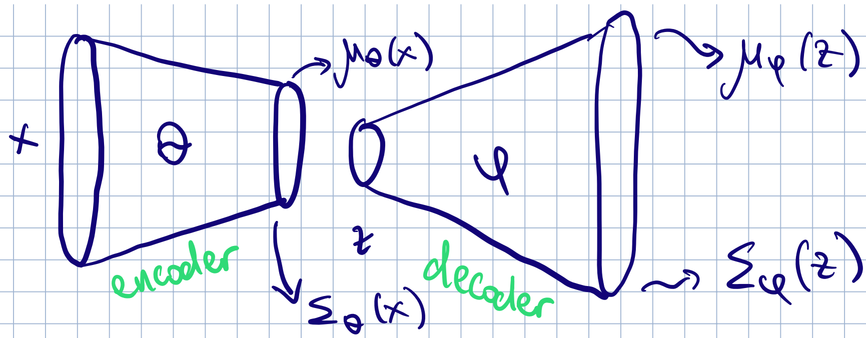
$$q = p(z|x,\varphi) \quad KL = 0$$

$$\mathcal{L}(q,\varphi) = \log p(x|\varphi)$$

q – arbitrary $\quad q(z) := q(z|x,\theta)$

$$\mathcal{L}(q,\varphi) = \int q(z|x,\theta) \log \frac{p(x,z|\varphi)}{q(z|x,\theta)}\,dz \to \max_{\varphi,\theta}$$

$$q(z|x,\theta) = \mathcal{N}(z|\mu_\theta(x), \Sigma_\theta(x))$$



$$\mu_\theta(x) \in \mathbb{R}^d \qquad \mu_\varphi(z) \in \mathbb{R}^D$$

$$\Sigma_\theta(x) = \mathrm{diag}\left(\sigma^2_{\theta 1}(x), \dots, \sigma^2_{\theta d}(x)\right) \qquad \Sigma_\varphi(z) = \mathrm{diag}\left(\sigma_{\varphi 1}(z), \dots, \sigma_{\varphi D}(z)\right)$$

$$\Sigma_\varphi(z) = I$$

$$\mathcal{L}(\theta, \varphi) = \int q(z \mid x, \theta) \log \frac{p(x, z \mid \varphi)}{q(z \mid x, \theta)} \, dz =$$

$$= \left\{ p(x, z \mid \varphi) = p(x \mid z, \varphi)\, p(z) \right\} =$$

$$= \int q(z \mid x, \theta) \log \frac{p(x \mid z, \varphi)\, p(z)}{q(z \mid x, \theta)} \, dz =$$

$$= \underbrace{\int q(z \mid x, \theta) \log p(x \mid z, \varphi)\, dz} + \underbrace{\int q(z \mid x, \theta) \log \frac{p(z)}{q(z \mid x, \theta)} \, dz}$$

$$\underbrace{\mathbb{E}_{q(z \mid x, \theta)} \log p(x \mid z, \varphi)}_{\text{(MSE) reconstruction loss}} - \underbrace{KL\left(q(z \mid x, \theta) \,\|\, p(z)\right)}_{\text{regularization}} \to \max_{\varphi, \theta}$$

$$\mathbb{E}_{q(z \mid x, \theta)} \log p(x \mid z, \varphi) \to D_\varphi, \; D_\theta$$

$$D_\varphi \, \mathbb{E}_{q(z \mid x, \theta)} \log p(x \mid z, \varphi)$$

$$\nabla_\theta \, \mathbb{E}_{q(z|x,\theta)} \log p(x|z,\varphi)$$

Reparametrization trick

$\xi \sim \mathcal{N}(\xi | m, s^2)$ $\quad$ $\varepsilon \sim \mathcal{N}(\varepsilon | 0, 1)$ $\quad$ $\xi = m + s \cdot \varepsilon$

$z \sim \mathcal{N}(z | \mu_\theta(x), \Sigma_\theta(x))$

$\varepsilon \sim \mathcal{N}(\varepsilon | 0, I)$ $\quad$ $\underbrace{z = \mu_\theta(x) + \Sigma_\theta^{1/2}(x) \cdot \varepsilon}$

$$\nabla_\theta \, \mathbb{E}_{q(z|x,\theta)} \log p(x|z,\varphi) = \nabla_\theta \, \mathbb{E}_{\varepsilon \sim \mathcal{N}(0,I)} \log p(x | \underbrace{z}_{z(\mu_\theta(x), \Sigma_\theta(x), \varepsilon)}, \varphi)$$

$$\log p(x|z,\varphi) = \log\left[\frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left(\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)\right] =$$

$$= -\frac{D}{2} \overset{\text{const}}{\cancel{\log 2\pi}} - \frac{1}{2} \log |\Sigma| + \frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) =$$

$$= -\frac{1}{2} \sum_{i=1}^{D} \log \sigma_i^2 + \frac{1}{2} \sum_{i=1}^{D} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \propto \{\Sigma_\varphi(z) = I\} \propto$$

$$\propto \frac{1}{2} \sum_{i=1}^{D} (\underset{\text{object}}{x_i} - \underset{\text{decoder output}}{\mu_i})^2$$

$q(z|x,\theta) = \mathcal{N}(z | \mu_\theta, \Sigma_\theta)$ $\quad$ $p(z) = \mathcal{N}(z | 0, I)$

$\xi_1 \sim \mathcal{N}(\xi_1 | \mu_1, \Sigma_1)$ $\quad$ $\xi_2 \sim \mathcal{N}(\xi_2 | \mu_2, \Sigma_2)$

$$KL(\xi_1 \| \xi_2) = \frac{1}{2}\left[\operatorname{tr}(\Sigma_2^{-1} \Sigma_1) - d + (\mu_2 - \mu_1) \Sigma_2^{-1} (\mu_2 - \mu_1) + \log \frac{|\Sigma_2|}{|\Sigma_1|}\right]$$

$$\Sigma_\theta = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_d^2)$$

$$KL(q \| p) = \frac{1}{2}\left[tr(\Sigma_\theta) - d + \|\mu_\theta\|_2^2 - \log|\Sigma_\theta| \right] =$$

$$= \frac{1}{2} \cdot \sum_{i=1}^{d} \left(\sigma_i^2 + \mu_i^2 - \log \sigma_i^2 \right) - \frac{d}{2}$$

$\log \sigma_i^2$ - output of linear layer

$\sigma_i^2 = \exp(\log \sigma_i^2)$