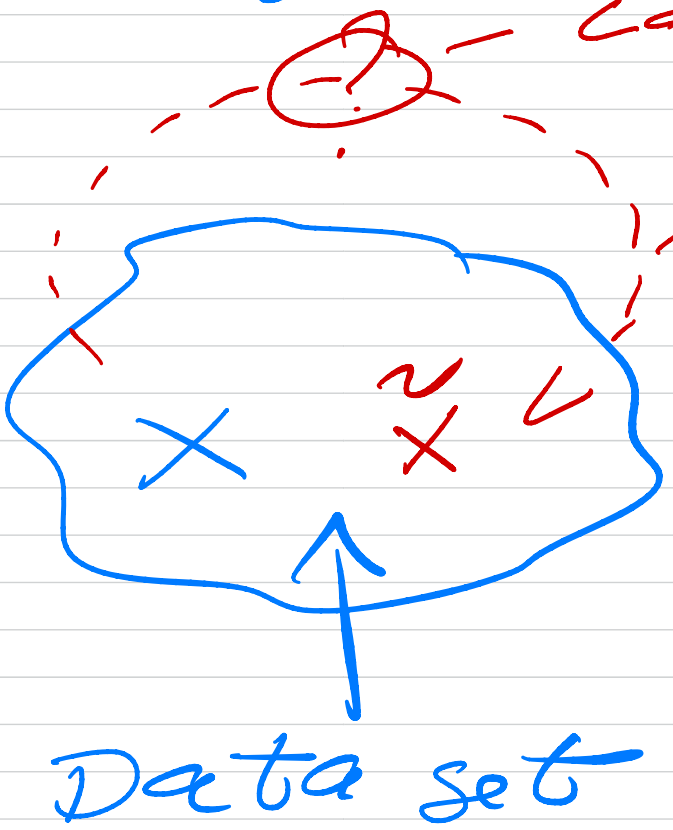
FYS5429/9429, April 10 lecture

FYS5429/9429 April 10

Basics of generative models



$q(z) \rightarrow p(z)$

$p(x;\theta)$

$p(x;\theta) =$

$\int dz\, p(z)\, p(x|z)$

We want $p(x;\theta)$

$X \in \underbrace{\{x_0, x_1, \ldots, x_{m-1}\}}_{D}$

$$p(X;\theta) = \prod_{x_i \in D} p(x_i;\theta)$$

(i.i.d.)

$$\underbrace{p(x_i;\theta)}_{\text{Marginal probability}} = \sum_{h_j} p(x_i, h_j;\theta)$$

hidden / latent

$$p(x_i;\theta) = \int dh\, p(x_i, h;\theta)$$

Boltzmann machines

$$p(x;\theta) = \frac{f(x;\theta)}{Z(\theta)}$$

$$Z(\theta) = \int_{x \in D} dx \int_h dh\, p(x, h;\theta)$$

normalization constant / partition function

$$p(X;\theta) = \frac{1}{Z(\theta)} \prod_{x_i \in X} f(x_i;\theta)$$

$$= \frac{1}{Z(\theta)} \prod_{x_i} \left[\sum_{h_j} f(x_i, h_j;\theta) \right]$$

optimize $\log p(X;\theta)$

$$\nabla_\theta \log p(X;\theta) = 0 =$$

$$\nabla_\theta \left[\sum_{x_i \in X} \log f(x_i;\theta) \right]$$

$$- \boxed{\left[E \left[\log f(x_i;\theta) \right] \right]} \quad - (MC)^2$$

RBM

$$a^T x + b^T h + x^T W h$$

$$f(x_i, h_j; \theta) = e$$

$$a^T x + b^T h + x^T W h =$$

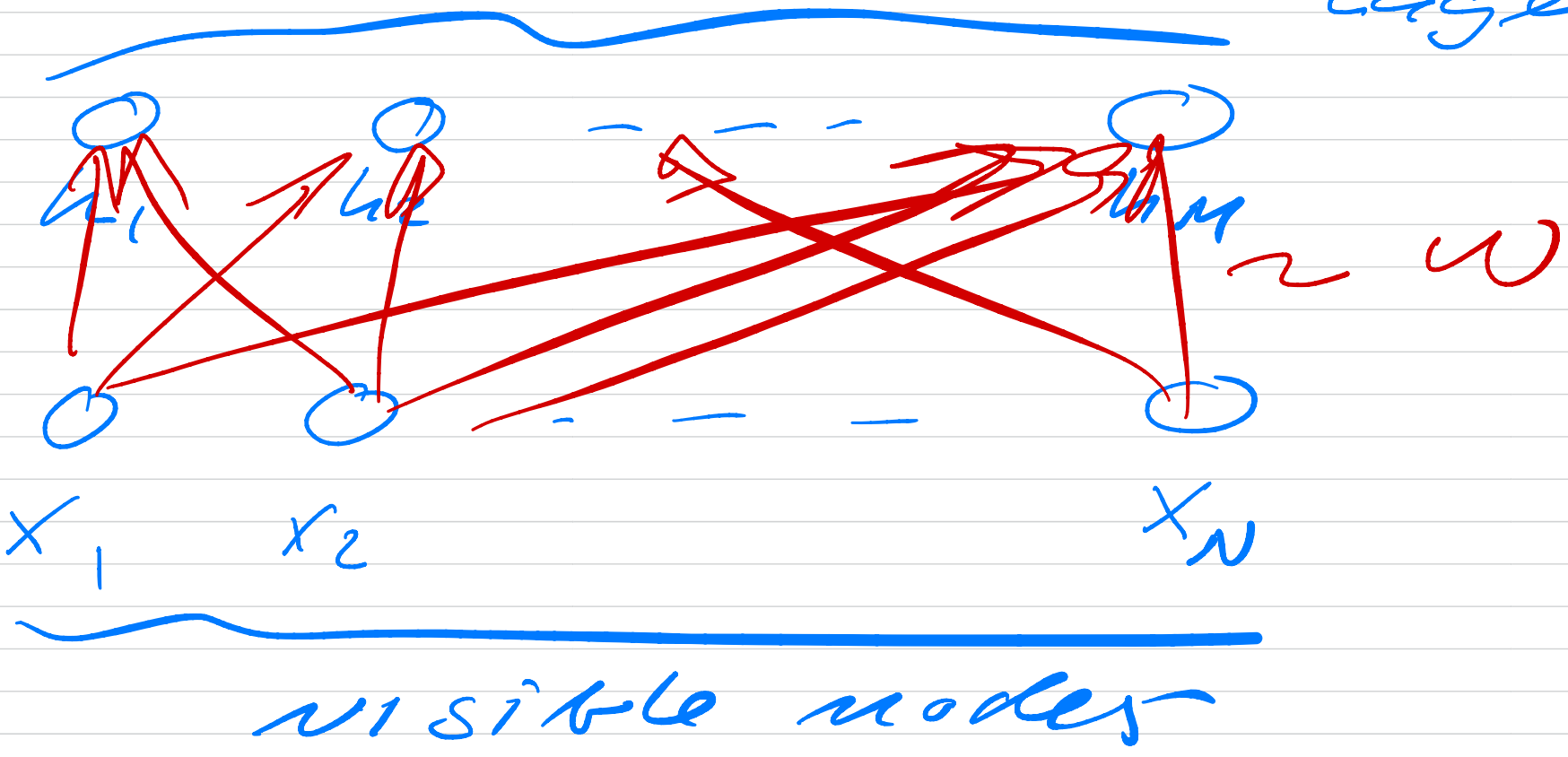
$$\sum_{i=1}^{N} a_i x_i + \sum_{j=1}^{M} b_j h_j + \sum_{i,j}^{MN} x_i w_{ij} h_j$$

Binary - Binary

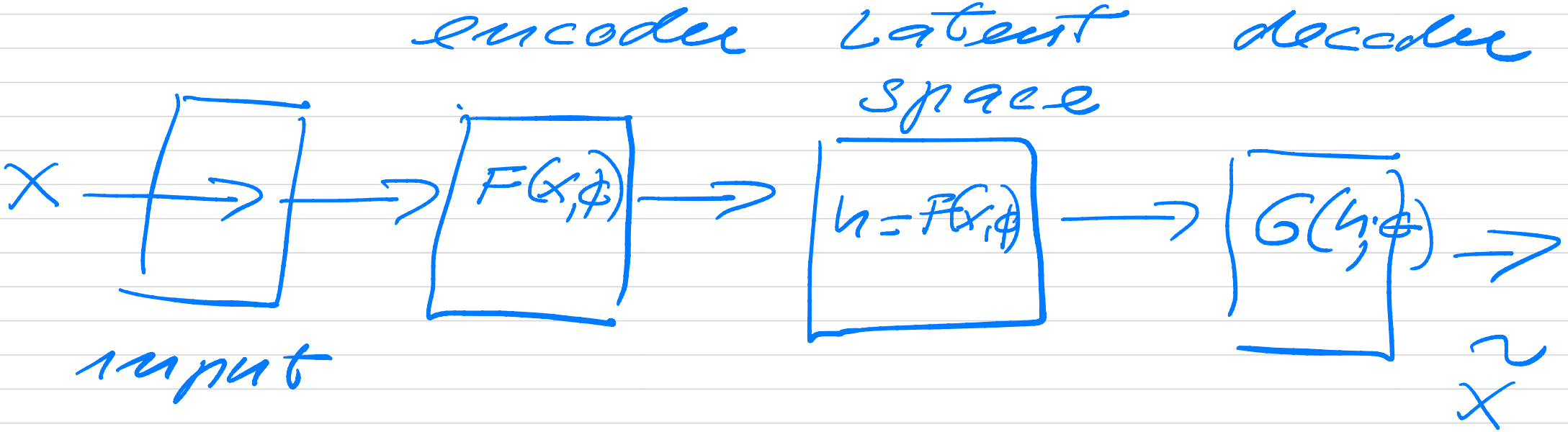
$$\theta = \{a, b, w\}$$

$$x_i = \{0, 1\} \qquad h_j = \{0, 1\}$$

NN for RBM



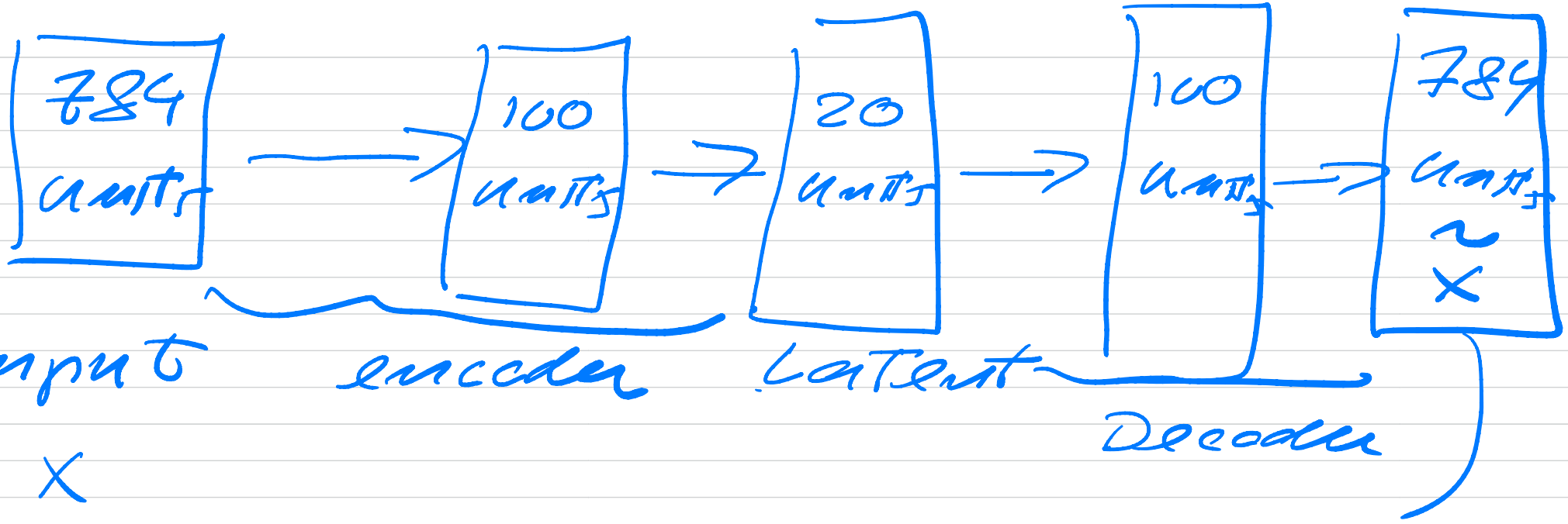
Reminder on Autoencoders



$$\tilde{x} = G(h;\theta)$$

reconstruction error

$$\hat{\theta}, \hat{\phi} = \underset{\theta,\phi}{\operatorname{argmin}} \|(x - \tilde{x})\|_2^2$$



Limitations of AEs

- Not generative, they don't define a distribution
- How to choose the latent dimension.

Why Variational AEs?

— A VAE is a generative model that learns a low-dimensional representation.

skip (θ, ϕ, etc)

conditional prob

$$p(x, h) = p(x|h)\,p(h) = p(h|x)\,p(x)$$

marginal prob

$$p(x) = \int dh\; p(x|h)\,p(h)$$

In case of discrete probs

$$p(x_i) = \sum_{h_j} p(x_i|h_j)\, p(h_j)$$

$$p(h_j) = \sum_{x_i} p(h_j|x_i)\, p(x_i)$$

Bayes' theorem

$$p(x|h) = \frac{p(h|x)\, p(x)}{\sum_{x_i} p(h_j|x_i)\, p(x_i)}$$

posterior — likelihood — prior

We want to optimize

$\log p(x)$ $\left(\log p(x;\theta)\right)$

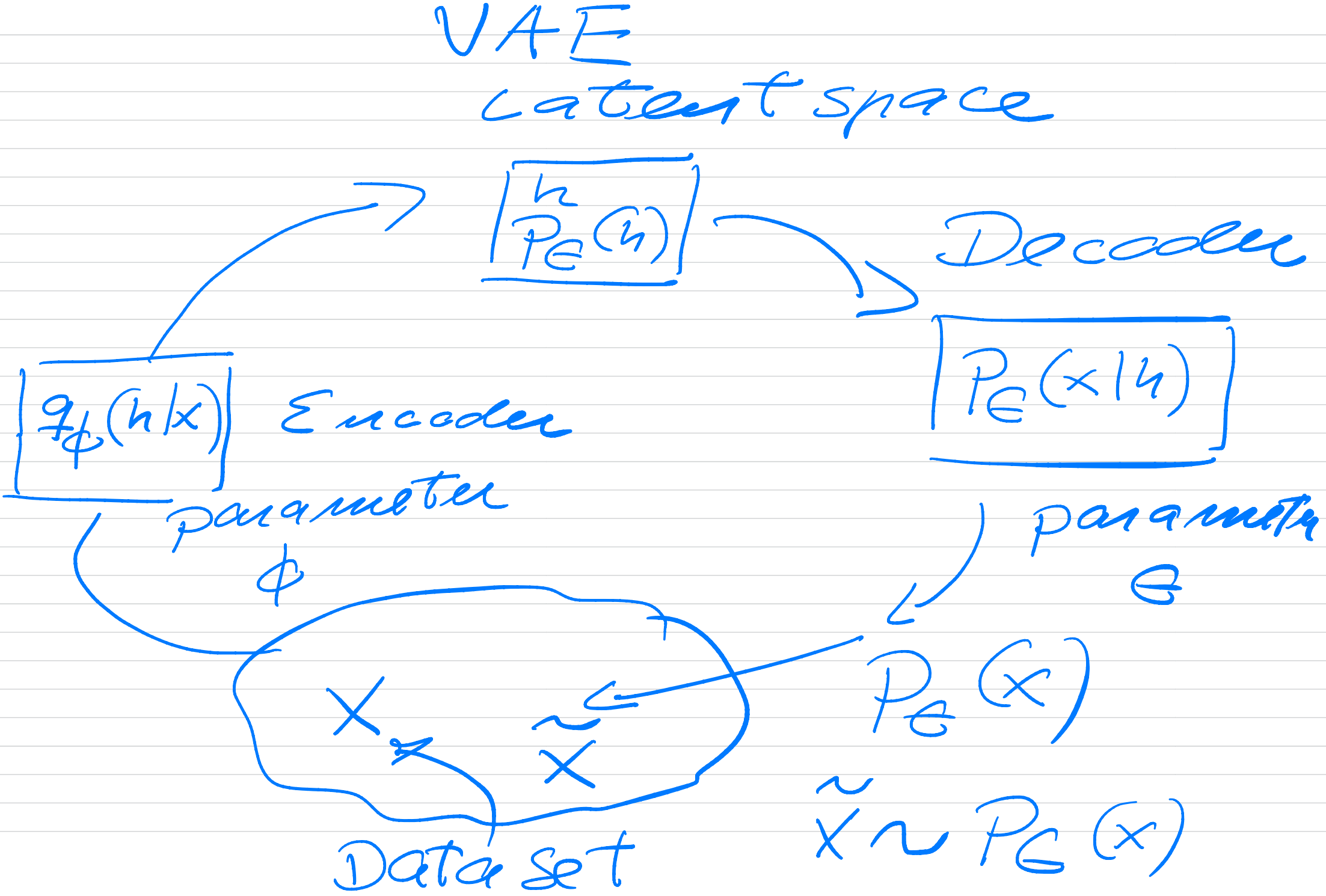
$= \log \left[\int p(h)\, p(x|h)\, dh \right]$

$= \log \int \frac{p(x,h)}{p(x|h)p(h)}\, dh$

insert $1 = \frac{q(h)}{q(h)}$

encoder $\rightarrow$ $q(h)$ latent space

Decoder


VAE
latent space
$p_\theta(h)$
Decoder
$p_\theta(x|h)$
$q_\phi(h|x)$
Encoder
parameter
ϕ
parameter
θ
$p_\theta(x)$
$\tilde{x} \sim p_\theta(x)$
Data set

$$\log p(x) = \log\left[\int p(h)\, p(x|h)\, dh\right]$$

q is probability

insert $1 = \frac{q(h)}{q(h)}$

Jensen's inequality

$$\geq \int q(h) \log\left[\frac{p(h)}{q(h)}\, p(x|h)\right] dh$$

$$= \int q(h) \log\left[\frac{p(h)}{q(h)}\right] dh +$$

$$\int q(h) \log(p(x|h))\, dh$$

$$= \underbrace{E_{h\sim q}\left[\log \frac{p(h)}{q(h)}\right]}_{(1)}$$

$$\left(h \sim q(h)\ (x \sim N(\mu_x, \nabla_x))\right)$$

$$+ \underbrace{E_{h\sim q}\left[\log p(x|h)\right]}_{(2)}$$

Let us look at the 2nd term first

assume

$$\log p(x|h) = \log \mathcal{N}(x; G_\theta(h), \sigma)$$

$$= \log \left[\frac{1}{(2\pi\sigma)^{d/2}} \exp\left[-\frac{1}{2\sigma^2} \|x - G_\theta(h)\|_2^2 \right] \right]$$

$$= -\frac{1}{2\sigma^2} \|x - G_\theta(h)\|_2^2 + \text{const}$$

(Think back to autoencoders)

$$\tilde{x} = G_\theta(h)$$

This is the reconstruction error

what about the first term?

- intermezzo: Kullback-Leibler divergence

$$KL(p\|q) = D_{KL}(p\|q)$$

$$= \int_{x \in D} p(x) \log \frac{p(x)}{q(x)} \, dx$$

$$= \mathbb{E}_{x \sim p}\left[\log \frac{p(x)}{q(x)} \right]$$

suppose $p(x)$ is the real data distribution and $q(x)$ is a model

$D_{KL}(p||q) = 0$ when $p(x) = q(x)$

when $p(x)$ tends to zero

$$\lim_{x \to 0^+} x \log x = 0$$

$$D_{KL}(p||q) \geq 0$$

$$\log p(x) = D_{KL}(q \| p)$$

— reconstruction error

For all:

$$\tilde{x} \sim p(x; \theta) = p_\theta(x)$$

want to optimize

$$\log p_\theta(x) = \log \left[\int p_\theta(x|h) p_\theta(h) \times dh \right]$$

a latent space whose distribution $q_\phi(h|x)$ links the input data with the latent space $-h-$

$$\hat{\theta}, \hat{\phi} = \underset{\theta, \phi}{\operatorname{argmax}} \log p_\theta(x)$$

$$= \log \underline{\underline{\left[\int p_\theta(x, h)\, dh \right]}}$$

$$= \log\left[\int du\, p_\theta(x|u)\, p_\theta(u)\right]$$

insert $1 = \dfrac{q_\phi(u|x)}{q_\phi(u|x)}$

$$= \log\left[\int \frac{p_\theta(x|u)}{q_\phi(u|x)}\, \frac{p_\theta(u)}{q_\phi(u|x)}\, du\right]$$

Jensen's inequality

$$\geq \mathbb{E}_{h \sim q_\phi}\left[\log p_\theta(x|h)\right]$$

$$(h \sim q_\phi(h|x))$$

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

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

Decoder

reconstruction error

$$- D_{KL}\left(q_\phi(u|x) \,\|\, p_\epsilon(u)\right)$$

$q_\phi(u|x)$ — encoder part

approximation

$$q_\phi(u|x) = N(u, \mu_\phi(x), \sigma_\phi(x))$$

multivariate

$$p_\epsilon(u) = N(u; 0, \mathbf{1})$$

mean

We can evaluate the D_{KL} analytically.

$$\arg\max_{\theta,\phi} \left\{ \mathbb{E}_{h\sim q_\phi}\left[\log p_\theta(x|h)\right] - D_{KL}\left[q_\phi(h|x)\,\|\,p_\theta(h)\right] \right\}$$

$\mathbb{E}_{h\sim q_\phi}\left[\log p_\theta(x|h)\right]$ → $(MC)^2$ - sampling

$D_{KL}\left[q_\phi(h|x)\,\|\,p_\theta(h)\right]$ → analytical evaluation