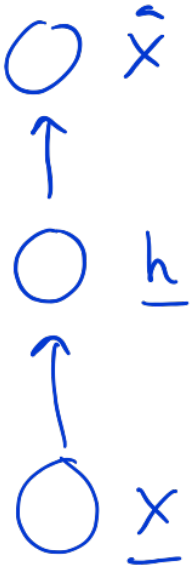
Representation Learning

Tirtharaj Dash

Dept. of CS & IS and APPCAIR
BITS Pilani, Goa Campus

November 27, 2021



Maximising $P(x \mid h)$

maximising $P(h \mid x)$

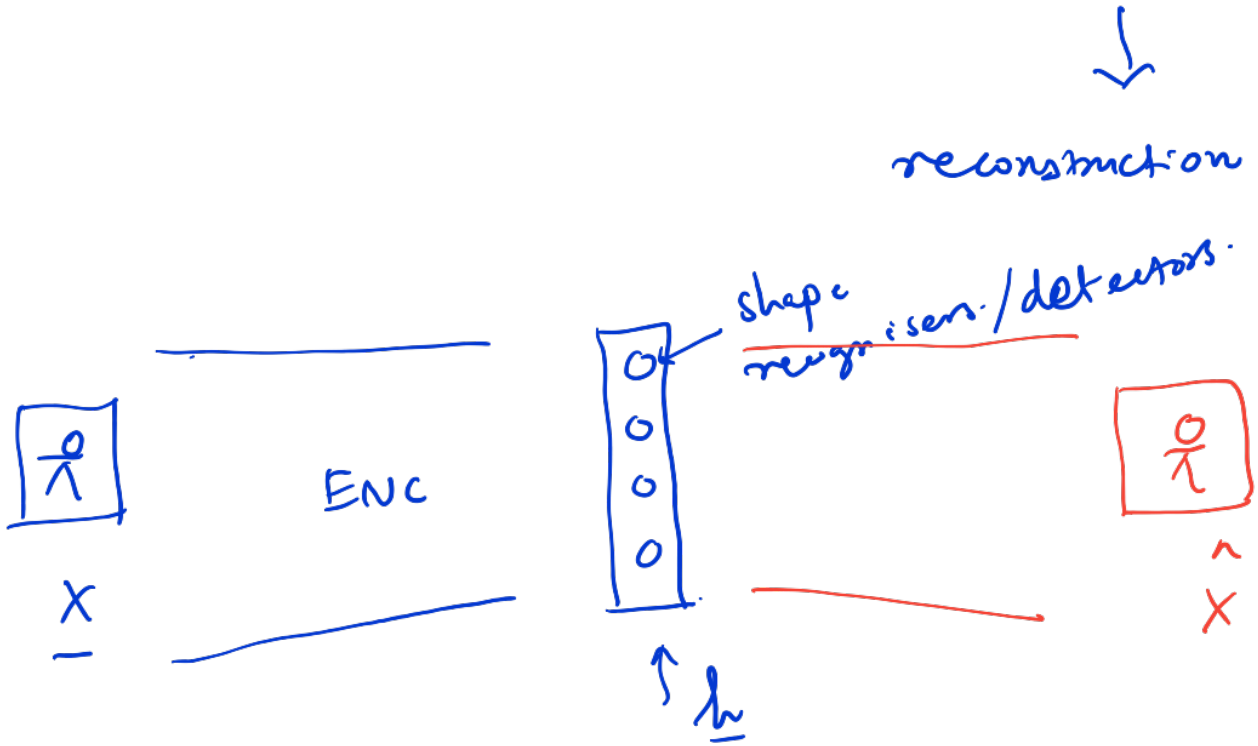
$\Big\}$ minimise the reconstruction loss

$L(x, \hat{x})$

$x \approx \hat{x}$

1. MSE
2. CE / NLL loss

Hinton " To recognise shapes, first learn to generate images. "

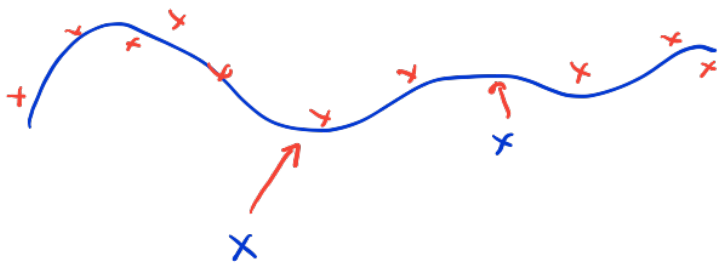


→ (1) sparse autoencoder.

$L + \lambda$ sparsity penalty → penalise h

→ (2) Robust to noise: Denoising autoencoder.

Denoising AE



x

$\downarrow$

$x + \text{Noise}$

$\downarrow$

$\sim N(0, \sigma^2)$

$\| $

λ^2

Regularisation for Representation learning:

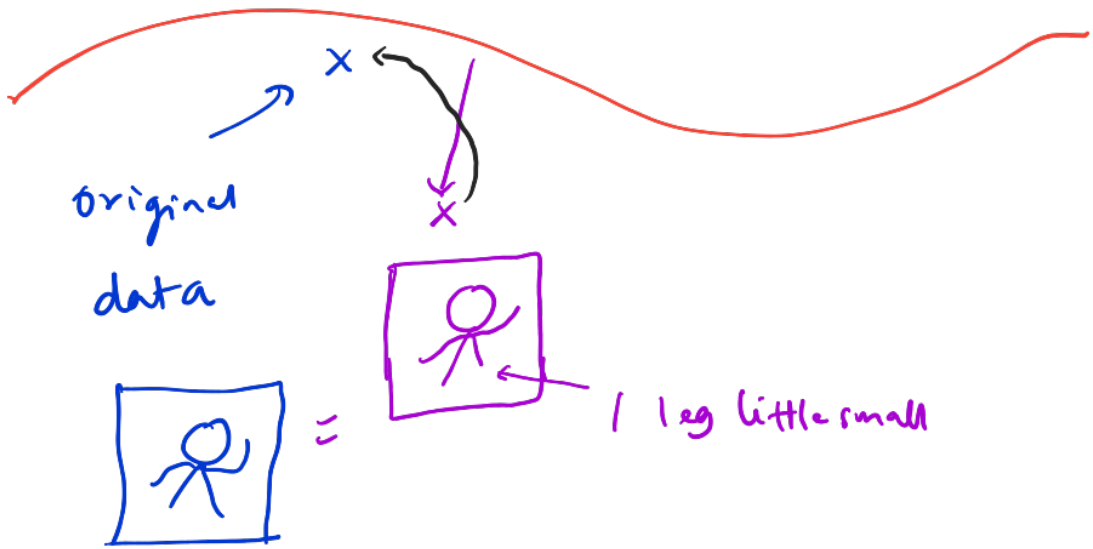
(1), (2)

(3) Contractive Autoencoder : Penalise the gradient.

Contractive AE:

→ Method to penalise a sparse representation.

→ Robustness to perturbation in data.



Let our reconstruction loss be:

$$L(\underline{x}, \underline{\hat{x}}) = \sum_{i=1}^{d} (x_i - \hat{x}_i)^2$$

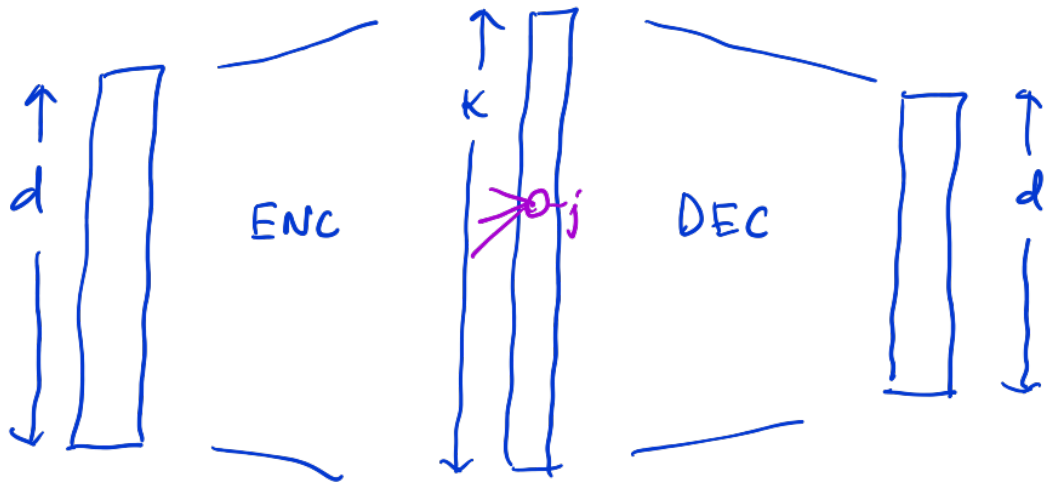
$\underline{x} \in \mathbb{R}^d$

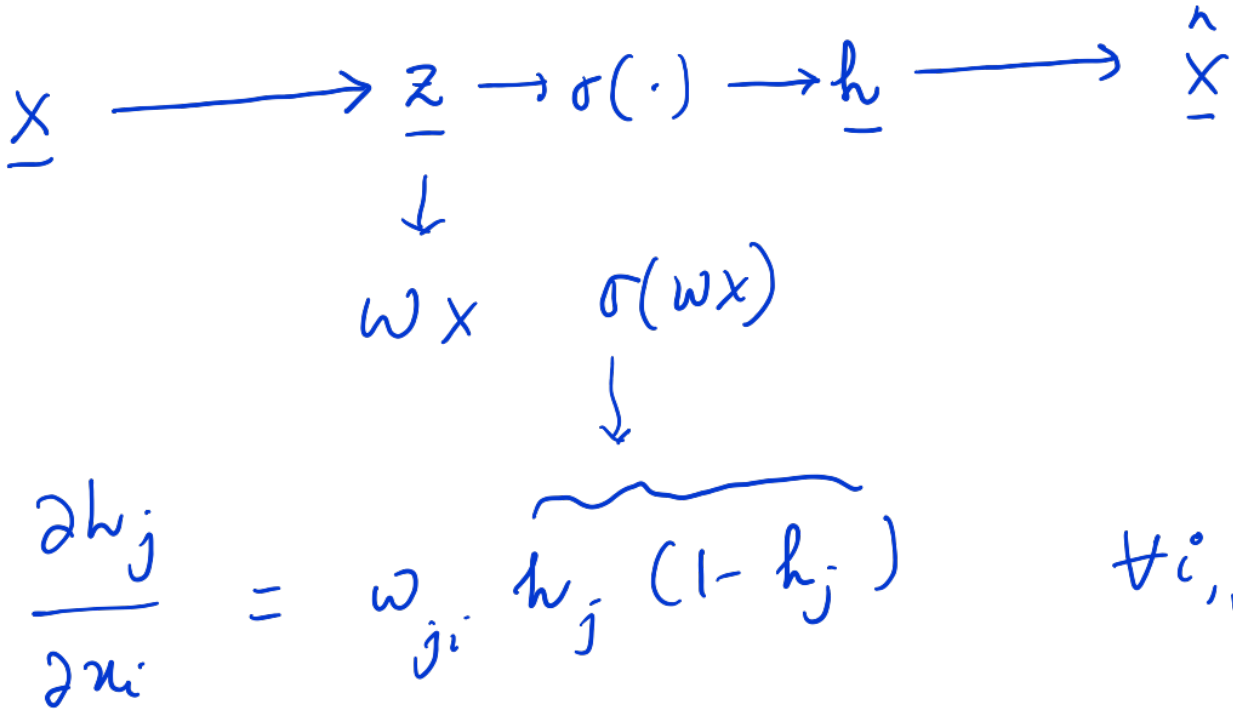
$\underline{\hat{x}} \in \mathbb{R}^d$

$\underline{h} \in \mathbb{R}^K$

In Contractive AE

$$R = \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{K} \left(\frac{\partial h_j}{\partial x_i} \right)^2$$





Total loss

$$L_{Total} = \sum_{i=1}^{d} (x_i - \hat{x}_i)^2 + \frac{\lambda}{2} \sum_{j=1}^{k} h_j^2 (1 - h_j)^2 \sum_{i=1}^{d} w_{ji}^2$$

after skipping some steps [HOMEWORK :-)]

$$\frac{\partial L_{Total}}{\partial w_{ji}} = x_i \frac{\partial L_{Total}}{\partial z_j} + \lambda w_{ji} h_j^2 (1 - h_j)^2$$

[VERIFY]

Task-specific loss

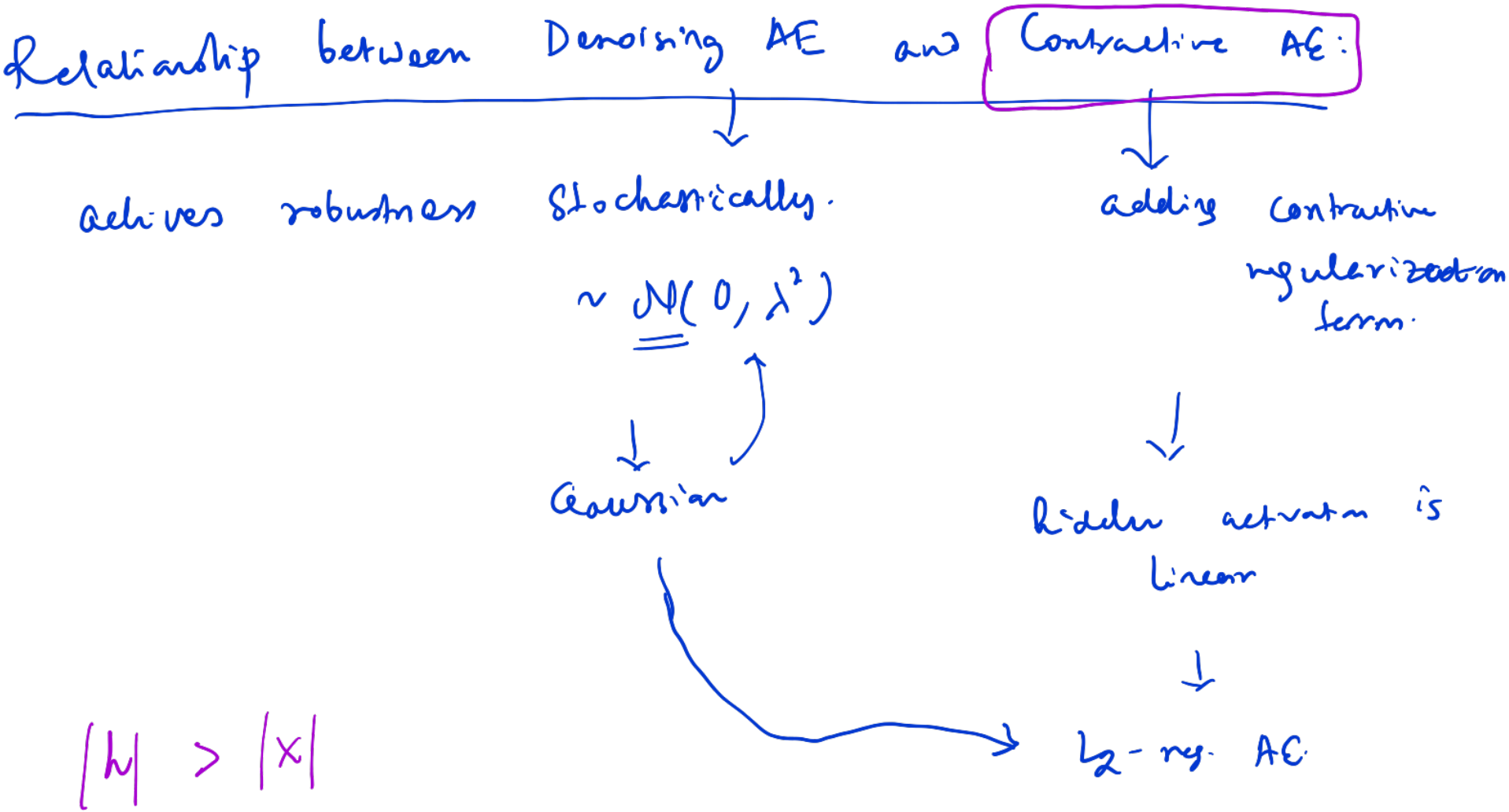
(reconstruction error should be $\to 0$)

Shrinking the hidden rep.

L_2-regularized AE

Q. Let hidden activation is linear?

$\lambda w_{ji} h_j^2$ $\quad \|h\|^2$


Relationship between Denoising AE and Contractive AE:
derives robustness
Stochastically.
$\sim \mathcal{N}(0, \lambda^2)$
Gaussian
adding contractive regularization term.
Hidden activation is Linear
L_2-reg. AE
$|h| > |x|$

Greedy-layerwise Pretraining using AE:

Without pretrain.

Minimizing loss.

↓

Maximum the likelihood

With Pretrain

$p(D|w)\,p(w)$

↓

Known data.

↓

Source

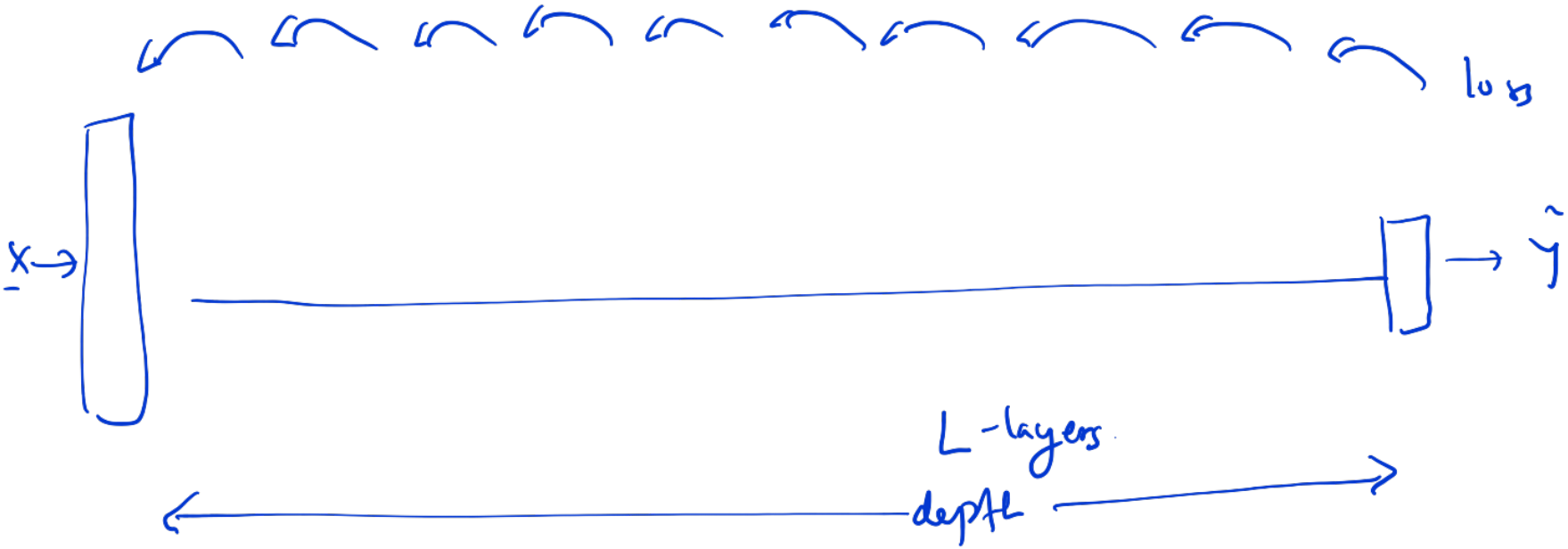
$p(h|D)$

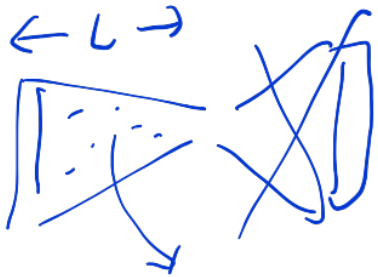
$\propto p(D|w)\,p(h)$

$\simeq$ $p(D|w)$ $p(w)$

likelihood. prior

↓



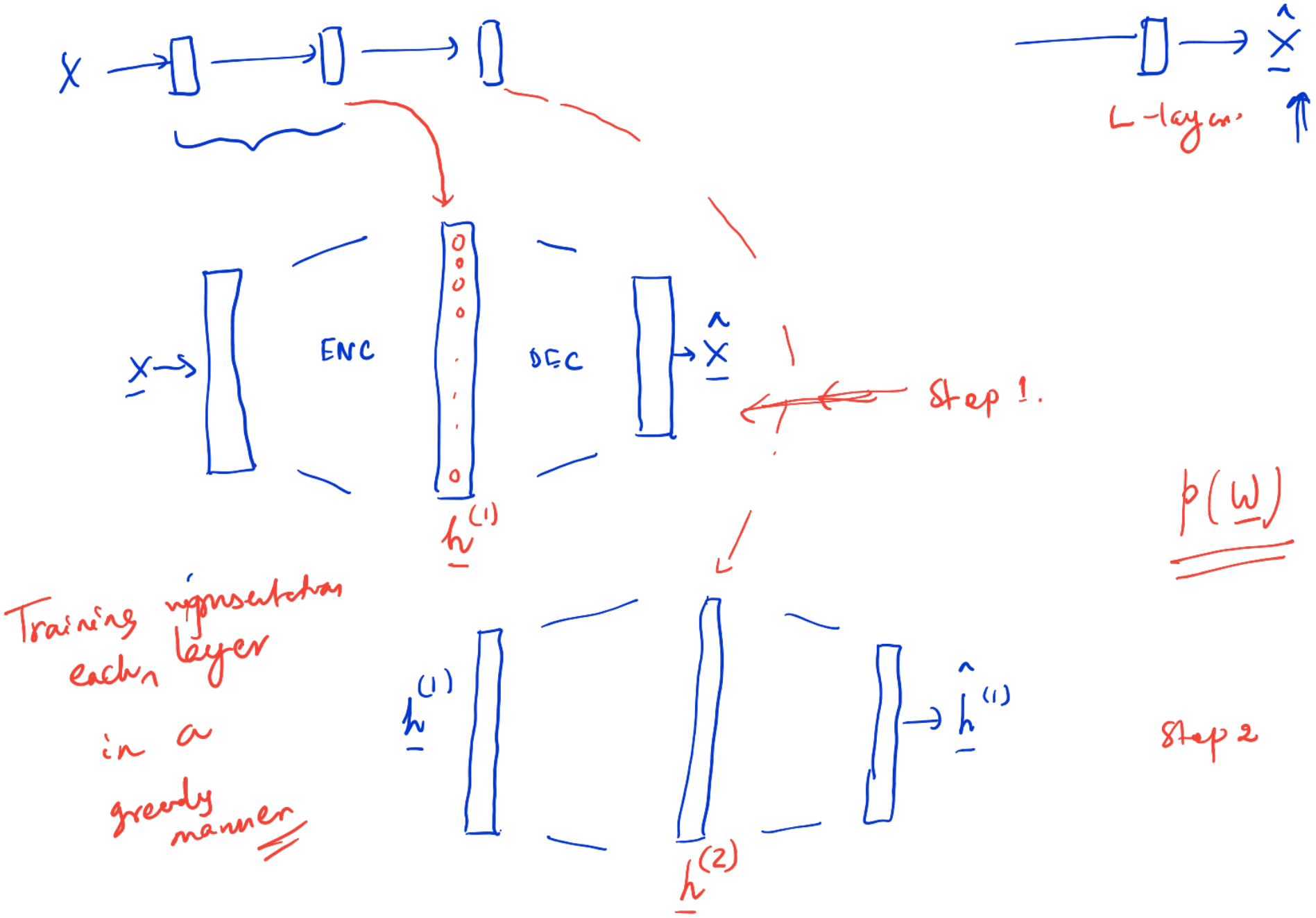


① resource

② computational issue

① Vanishing grad.

⋮


X
L-layer.
ENC
DEC
Step 1.
p(W)
Training unsupervised each layer in a greedy manner
Step 2

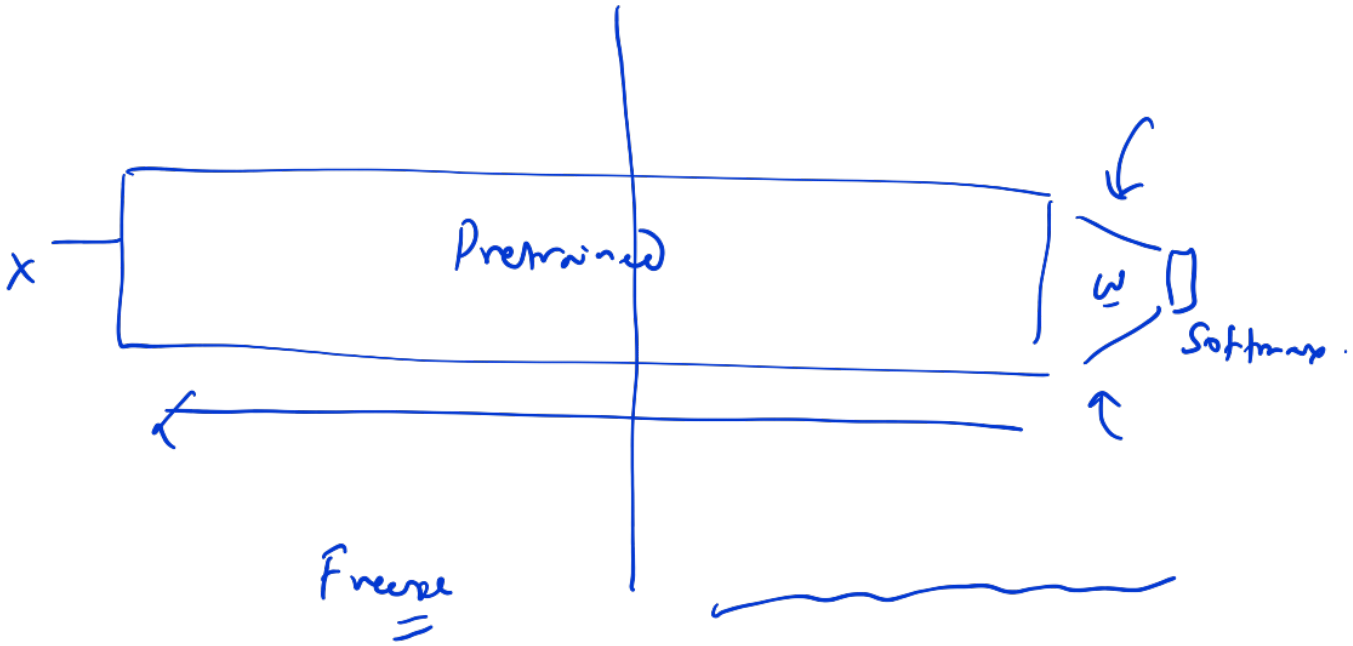
Next Class:

Learning "hidden" / "latent" Probabilistic structure from data.

(VAE)

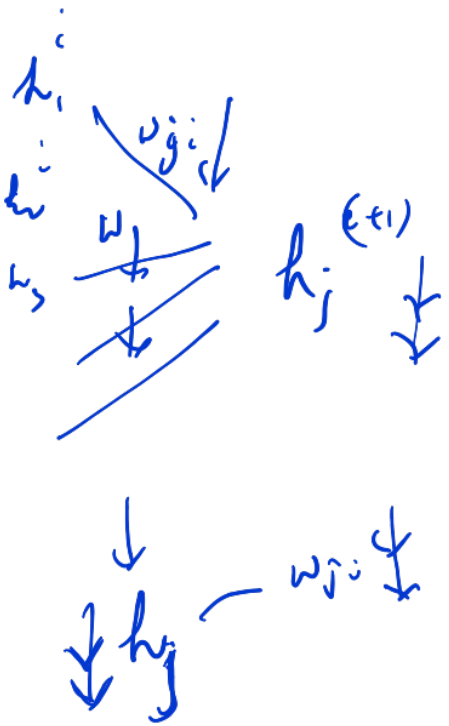
(GAN)

$p(\underline{x})$


x
Pretrained
w
Softmax
Freeze

$$\frac{\partial L_{Total}}{\partial w_{ji}} = \square + \lambda w_{ji} \, h_j^2 (1-h_j)^2$$

$$w_{ji} \leftarrow w_{ji} - \eta \left[\frac{\partial L_{Total}}{\partial w_{ji}} \right]$$



$$\leftarrow w_{ji} - \eta [\quad] \quad h(1-h)$$

$$- \eta \left[\lambda w_{ji} \, h_j^2 (1-h_j)^2 \right]$$

large