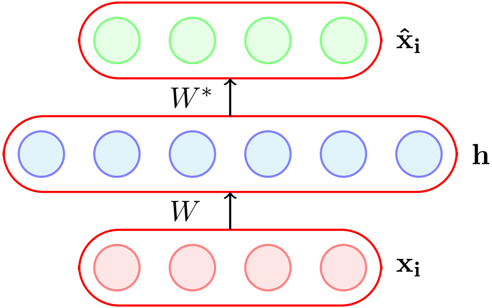
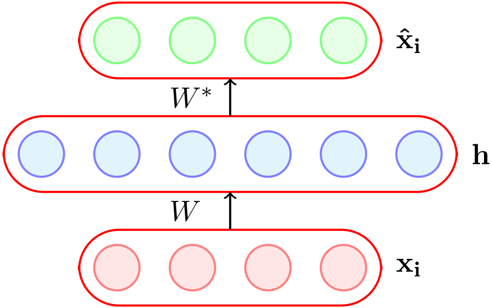
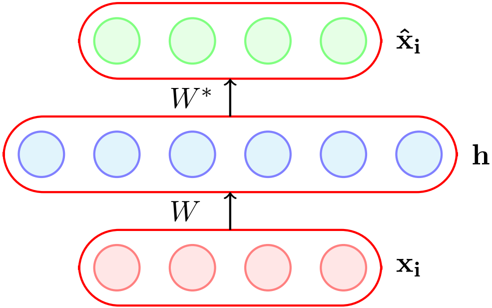
Module 7.3: Regularization in autoencoders (Motivation)

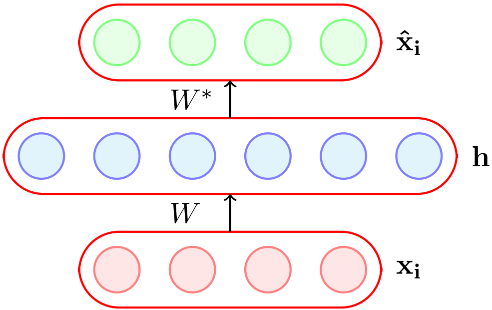

$\hat{\mathbf{x}}_\mathbf{i}$
W^*
$\mathbf{h}$
W
$\mathbf{x}_\mathbf{i}$



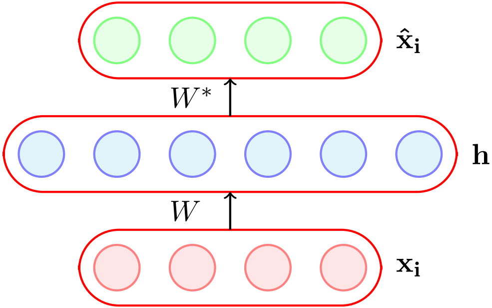
- While poor generalization could happen even in undercomplete autoencoders it is an even more serious problem for overcomplete auto encoders



- While poor generalization could happen even in undercomplete autoencoders it is an even more serious problem for overcomplete auto encoders
- Here, (as stated earlier) the model can simply learn to copy $\mathbf{x}_\mathbf{i}$ to $\mathbf{h}$ and then $\mathbf{h}$ to $\hat{\mathbf{x}}_\mathbf{i}$

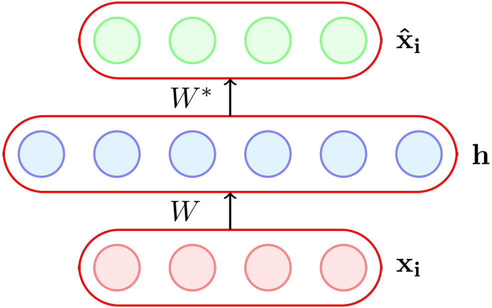


- While poor generalization could happen even in undercomplete autoencoders it is an even more serious problem for overcomplete auto encoders
- Here, (as stated earlier) the model can simply learn to copy $\mathbf{x}_\mathbf{i}$ to $\mathbf{h}$ and then $\mathbf{h}$ to $\hat{\mathbf{x}}_\mathbf{i}$
- To avoid poor generalization, we need to introduce regularization



- The simplest solution is to add a L_2-regularization term to the objective function

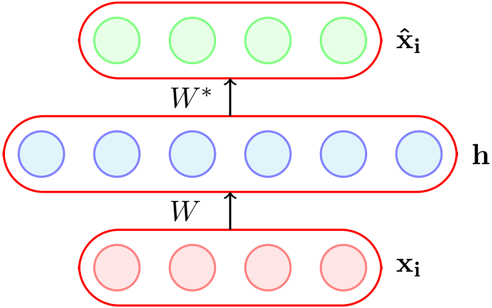
$$\min_{\theta, w, w^*, \mathbf{b}, \mathbf{c}} \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{x}_{ij} - x_{ij})^2 + \lambda \|\theta\|^2$$



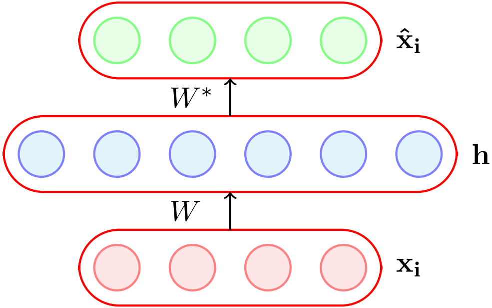
- The simplest solution is to add a L_2-regularization term to the objective function

$$\min_{\theta,w,w^*,\mathbf{b},\mathbf{c}} \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}(\hat{x}_{ij}-x_{ij})^2+\lambda\|\theta\|^2$$

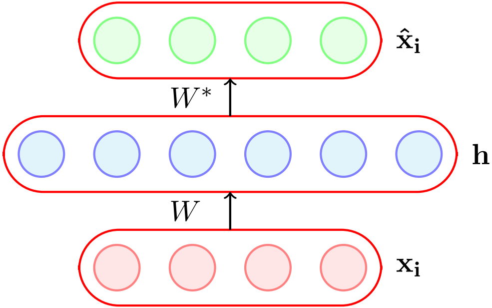
- This is very easy to implement and just adds a term λW to the gradient $\frac{\partial \mathscr{L}(\theta)}{\partial W}$ (and similarly for other parameters)



- Another trick is to tie the weights of the encoder and decoder



- Another trick is to tie the weights of the encoder and decoder i.e., $W^* = W^T$



- Another trick is to tie the weights of the encoder and decoder i.e., $W^* = W^T$
- This effectively reduces the capacity of Autoencoder and acts as a regularizer