Deep Learning

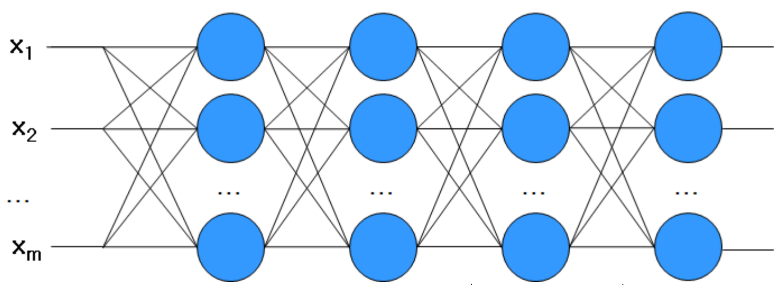
Content

- Vanishing Gradient & Activation Functions
- Dropout
- Batch Normalization

Gradient Vanishing & Activation Functions

Gradient Vanishing & Exploding

- **Gradient is easy to vanish or explode**
  - To many terms are multiplied.
  - If some are small numbers, gradient becomes very small.
  - If some are large numbers, gradient becomes very large.



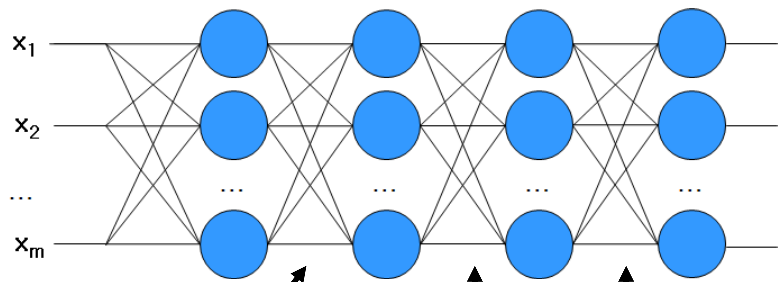
$$\frac{\partial E_n}{\partial w_{kj}} = -(t_{nk} - o_{nk})o_{nk}(1 - o_{nk})h_{nj}$$

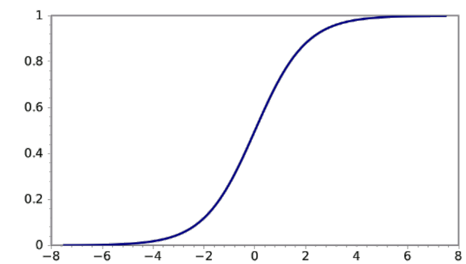
$$\frac{\partial E_n}{\partial w_{ji}} = -h_{nj}(1 - h_{nj})x_{ni}\sum_{k=1}^{m} w_{kj}(t_{nk} - o_{nk})o_{nk}(1 - o_{nk})$$

$$\frac{\partial E}{\partial w_{ip}} = \left(\sum_{j=1}^{J}\left(\sum_{k=1}^{K} -(t_n - o_{nk})o_{nk}(1 - o_{nk})w_{kj}\right)h_{nj}(1 - h_{nj})w_{ji}\right)h_{ni}(1 - h_{ni})h_{np}$$

Activation Function

- **Vanishing Gradient**
  - The major terms are the derivatives of the activation function





$$\frac{\partial \mathbf{Sigmoid}}{\partial w} \le \frac{1}{4}$$

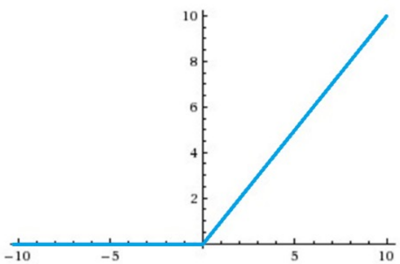
$$\frac{\partial E_n}{\partial w_{kj}} = -(t_{nk} - o_{nk})o_{nk}(1 - o_{nk})h_{nj}$$

$$\frac{\partial E_n}{\partial w_{ji}} = -h_{nj}(1 - h_{nj})x_{ni}\sum_{k=1}^{m} w_{kj}(t_{nk} - o_{nk})o_{nk}(1 - o_{nk})$$

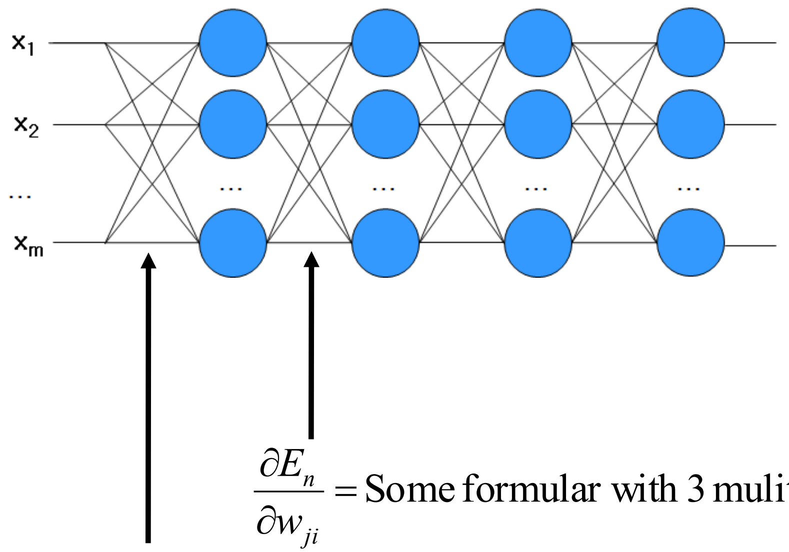
$$\frac{\partial E}{\partial w_{ip}} = \left(\sum_{j=1}^{J}\left(\sum_{k=1}^{K} -(t_n - o_{nk})o_{nk}(1 - o_{nk})w_{kj}\right)h_{nj}(1 - h_{nj})w_{ji}\right)h_{ni}(1 - h_{ni})h_{np}$$

Activation Function

- **Using another functions instead of sigmoid**
  - Rectified Linear Unit (ReLU)



$$f(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$



$$\frac{\partial E_n}{\partial w_{ji}} = \text{Some formular with 3 mulitiplications of } \frac{\partial f}{\partial w}$$

$$\frac{\partial E_n}{\partial w_{hg}} = \text{Some formular with 4 mulitiplications of } \frac{\partial f}{\partial w}$$

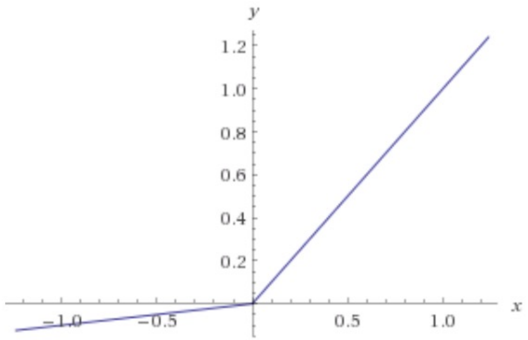
Activation Function

- **Advantage**
  - No vanishing gradient problems.
    - Deep networks can be trained without pre-training
  - Sparse activation
    - In a randomly initialized network, only about 50% of hidden units are activated
  - Fast computation:
    - 6 times faster than sigmoid function
- **Disadvantage**
  - Knockout Problem

Activation Function

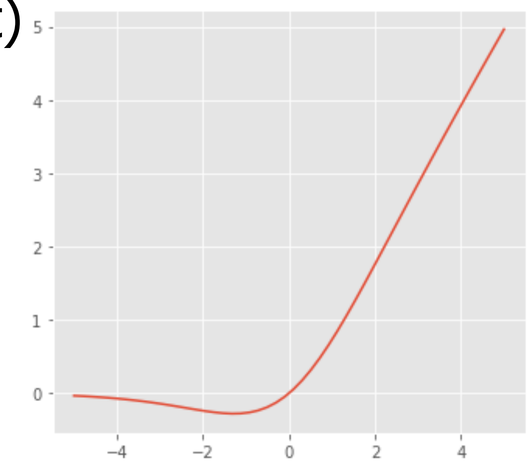
- **You may use another**
  - Leaky ReLU

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ 0.01x & \text{otherwise} \end{cases}$$

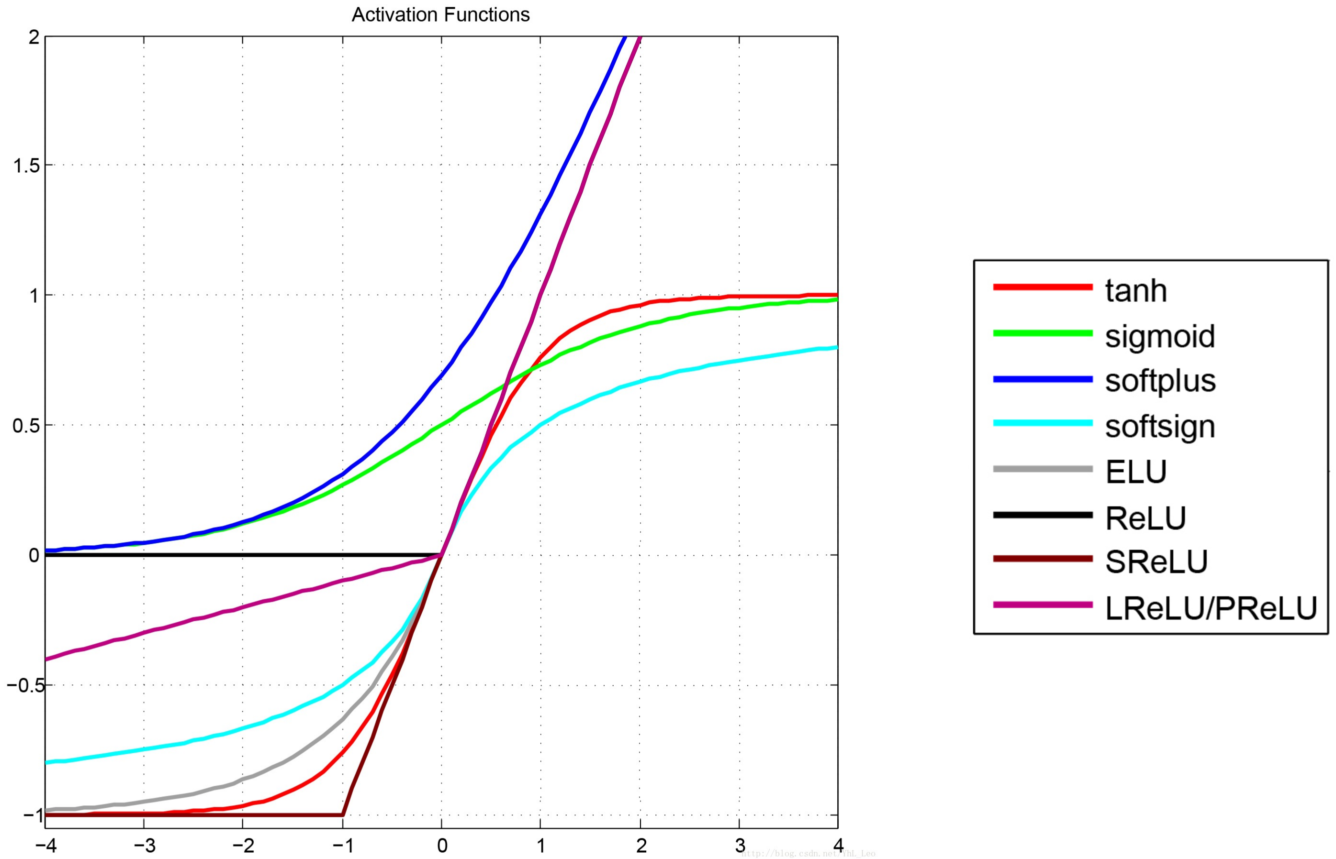


  - Swish (or SiLU-Sigmoid Linear Unit)

$$f(x) = \frac{x}{1 + e^{-x}}$$



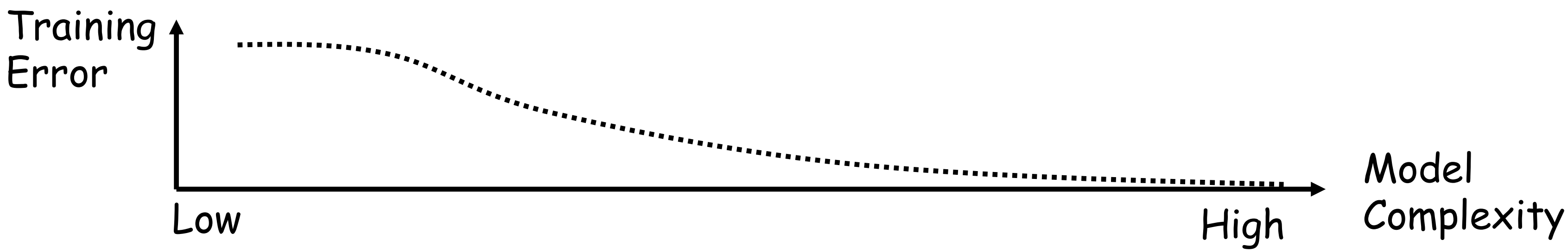
Other Activation Functions

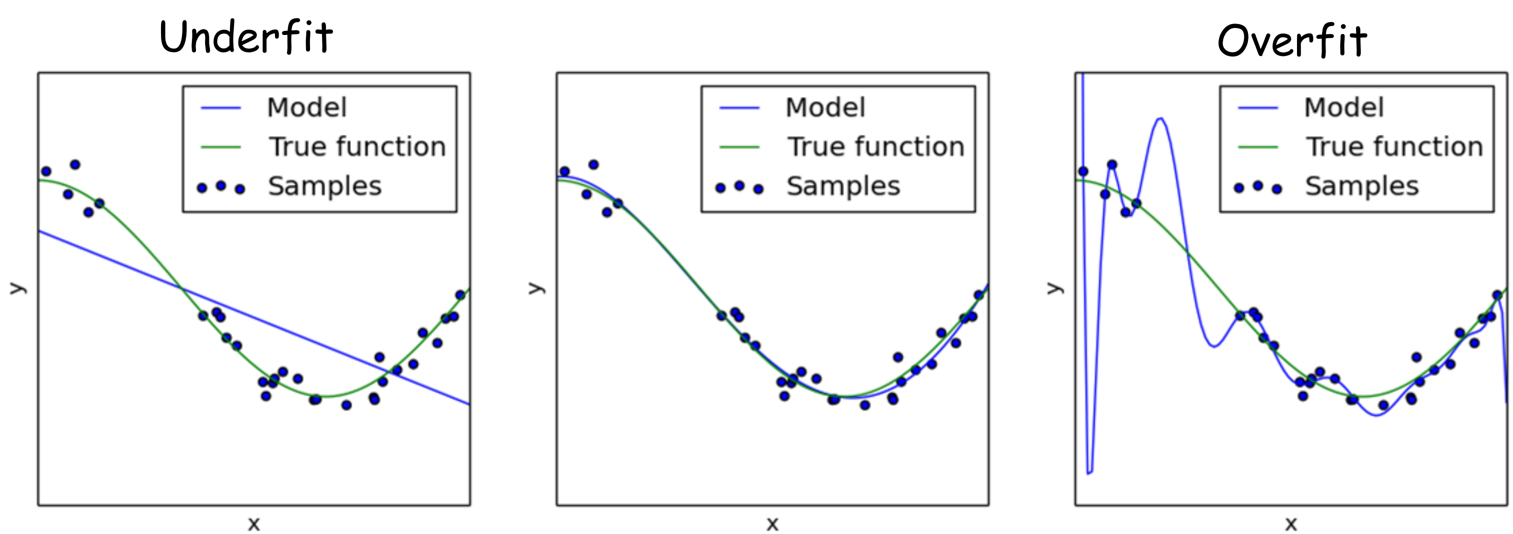


Regularization

Overfitting

- **Overfitting**





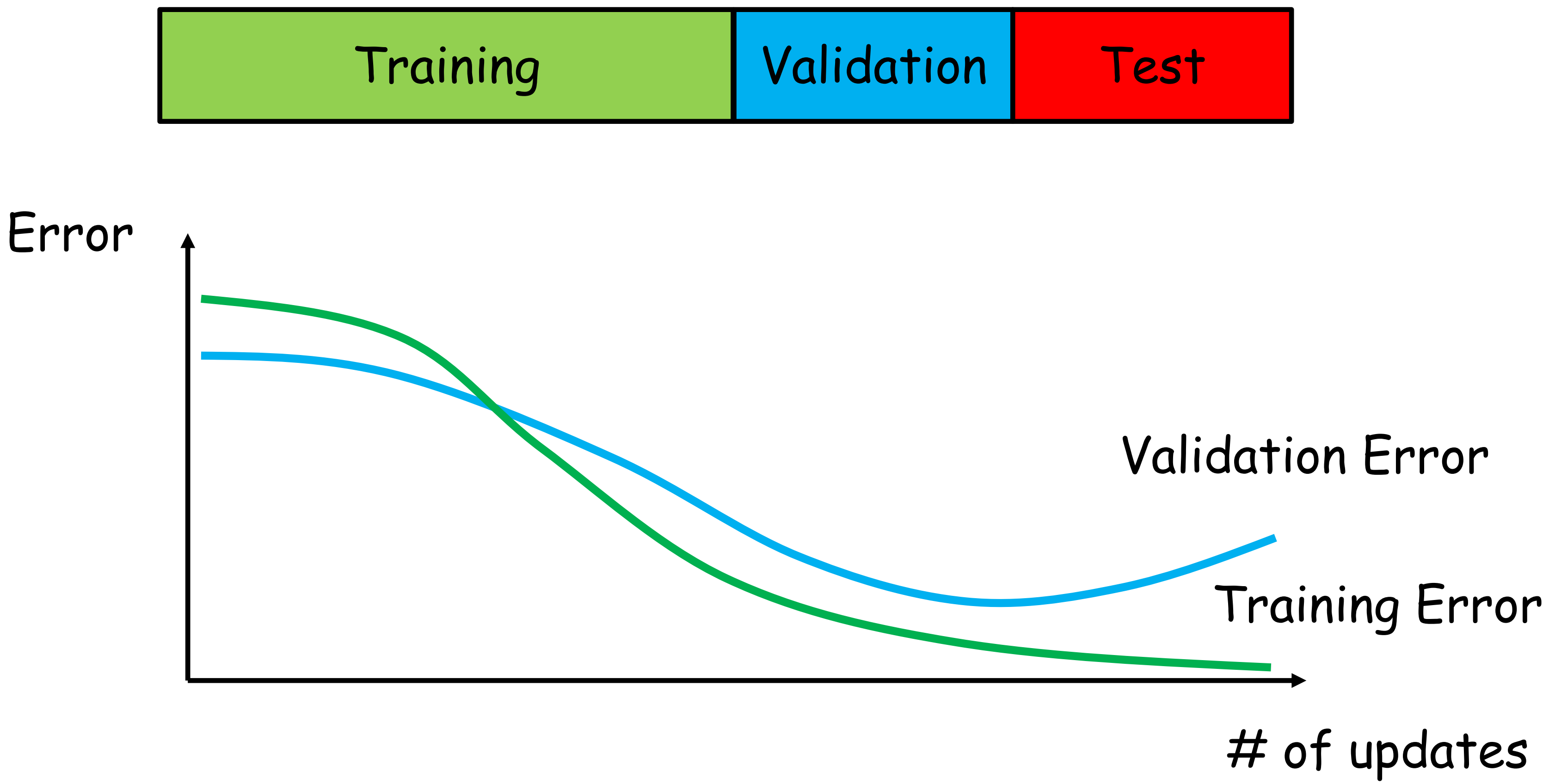
Regularization

- **What is Regularization**
  - Introducing additional information to prevent over-fitting

- **Approaches**
  - Proper Learning: Early stopping
  - Proper Structure: Weight decay, Dropout, DropConnect, Stochastic pooling

Early Stopping

- **Split data into 3 groups**



Weight Decay

- **L1 Regularization**
  - Leading most weights very close to zero
  - Choosing a small subset of most important inputs
  - Resistant to noise in the inputs.

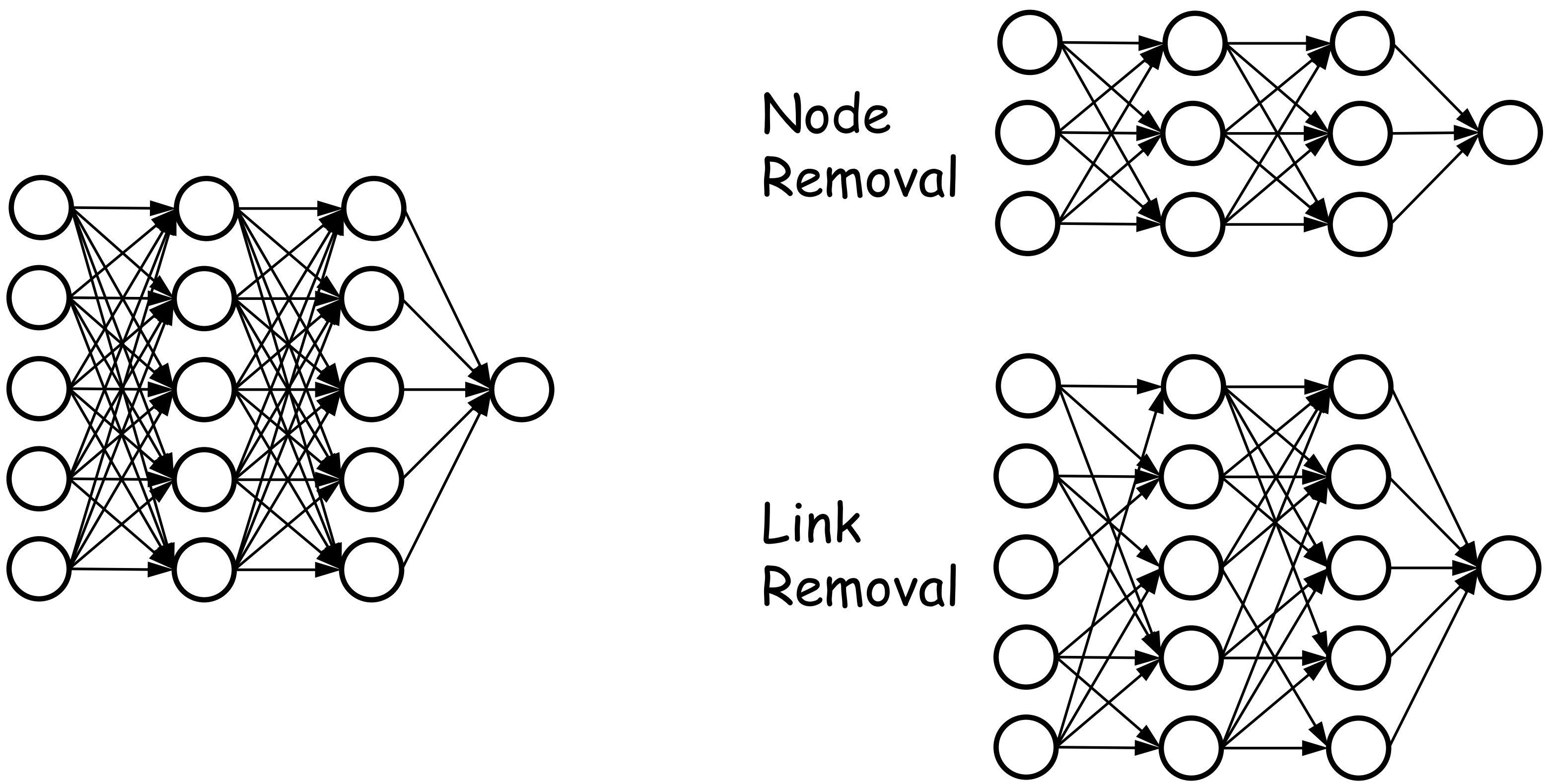
$$\widetilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\lambda}{2}\left|\mathbf{w}\right|$$

- **L2 Regularization**
  - Penalizing peaky weights
  - Encouraging to use all of its inputs a little rather than using only some of its inputs a lot.

$$\widetilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

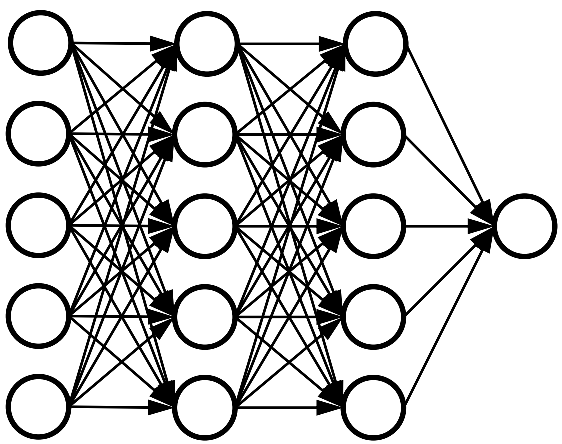
Weight Decay

- **Complex Structure vs Simple Structure**

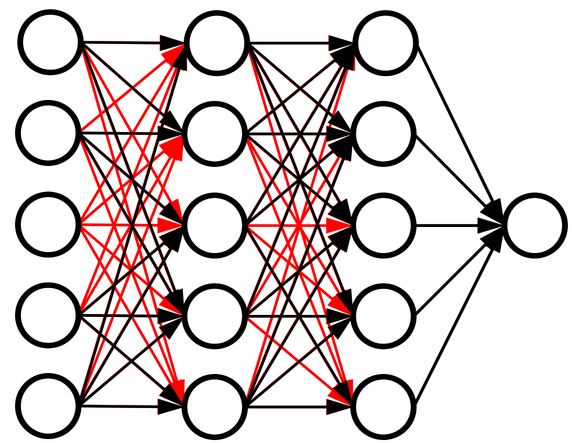


Weight Decay

- **Complex Structure vs Simple Structure**



Set
many links
to zero

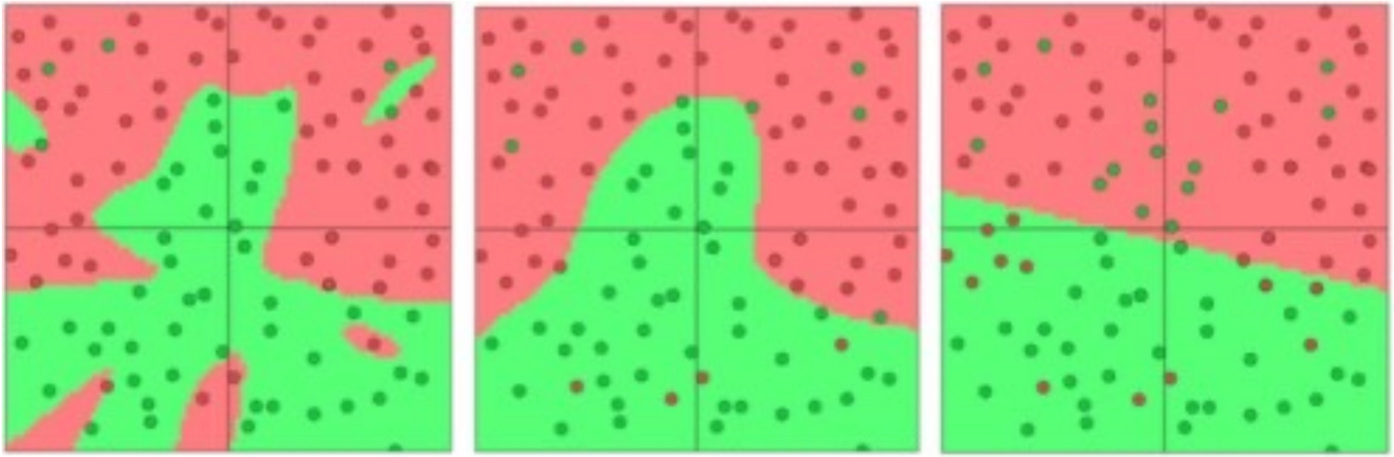


|w| is large <-> NN is Complex

|w| is small <-> NN is Simple

Weight Decay

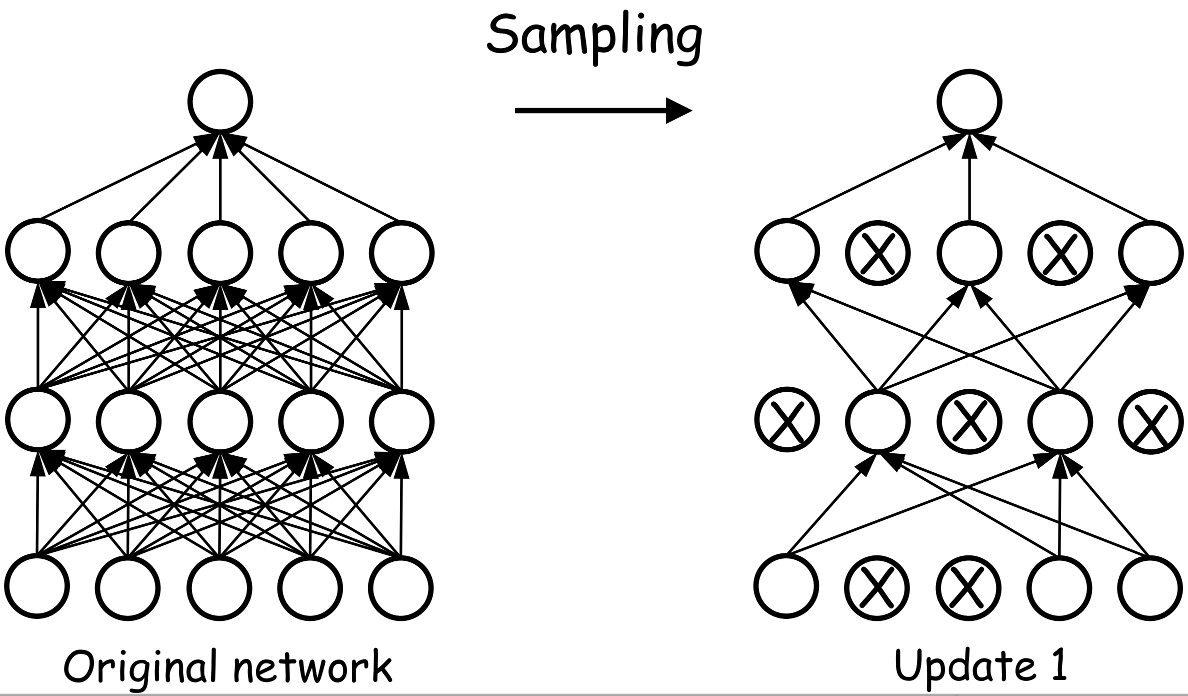
- **Example: Separating green and red**



L2 regularization strengths of 0.01, 0.1, and 1

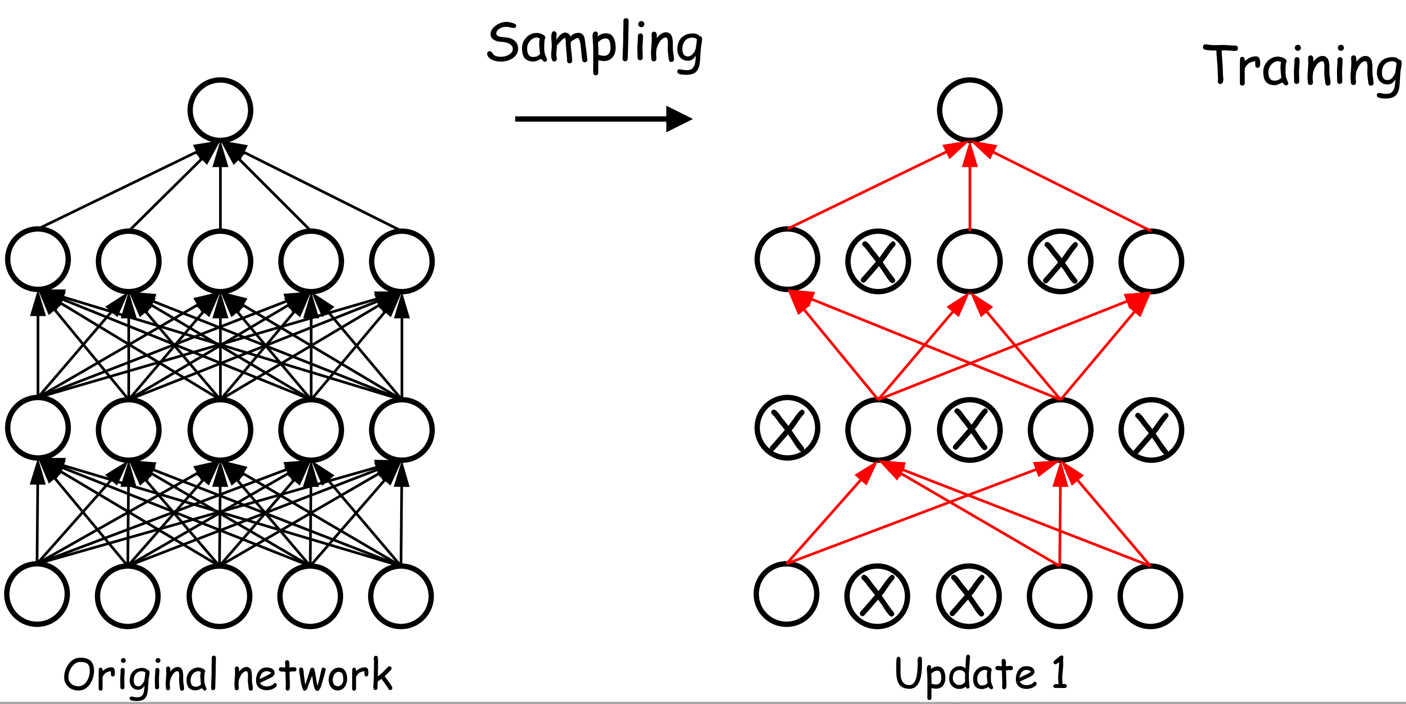
Dropout

- **How can we reduce the structural complexity without removing nodes?**
  - Hmm??



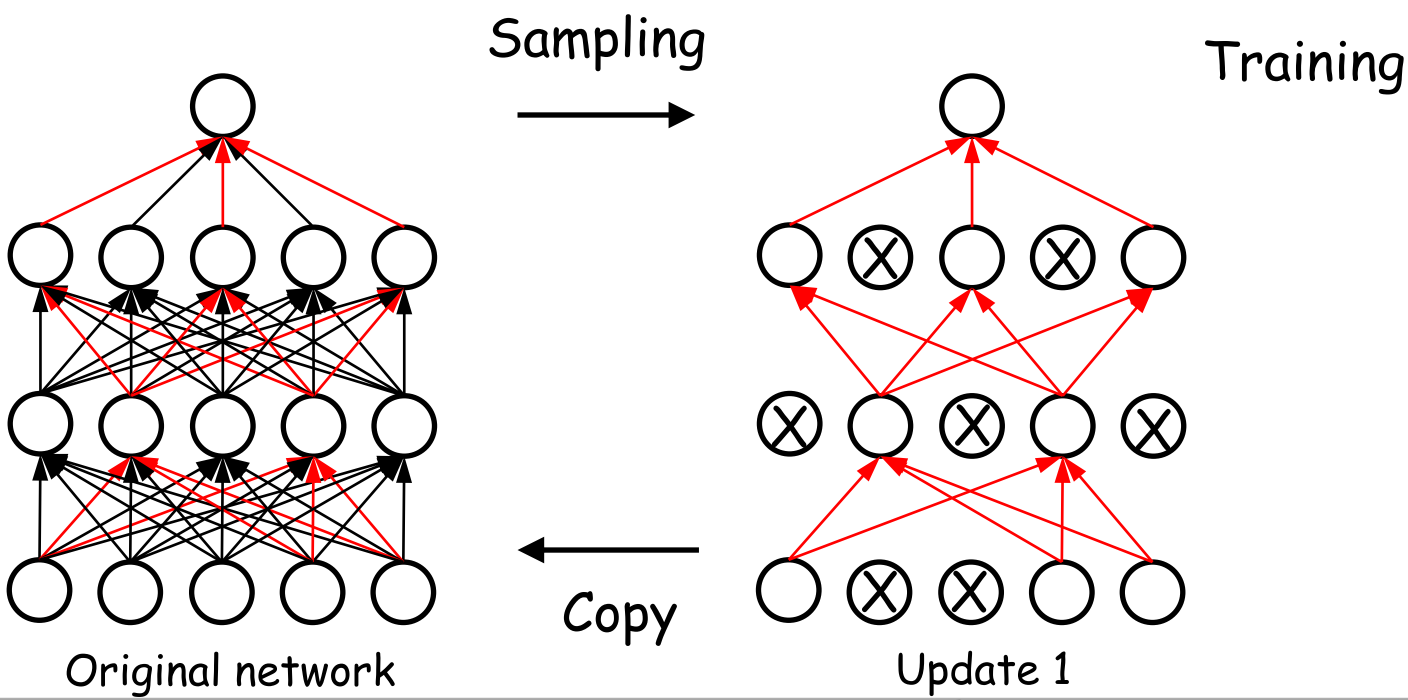
Dropout

- **How can we reduce the structural complexity without removing nodes?**
  - Hmm??



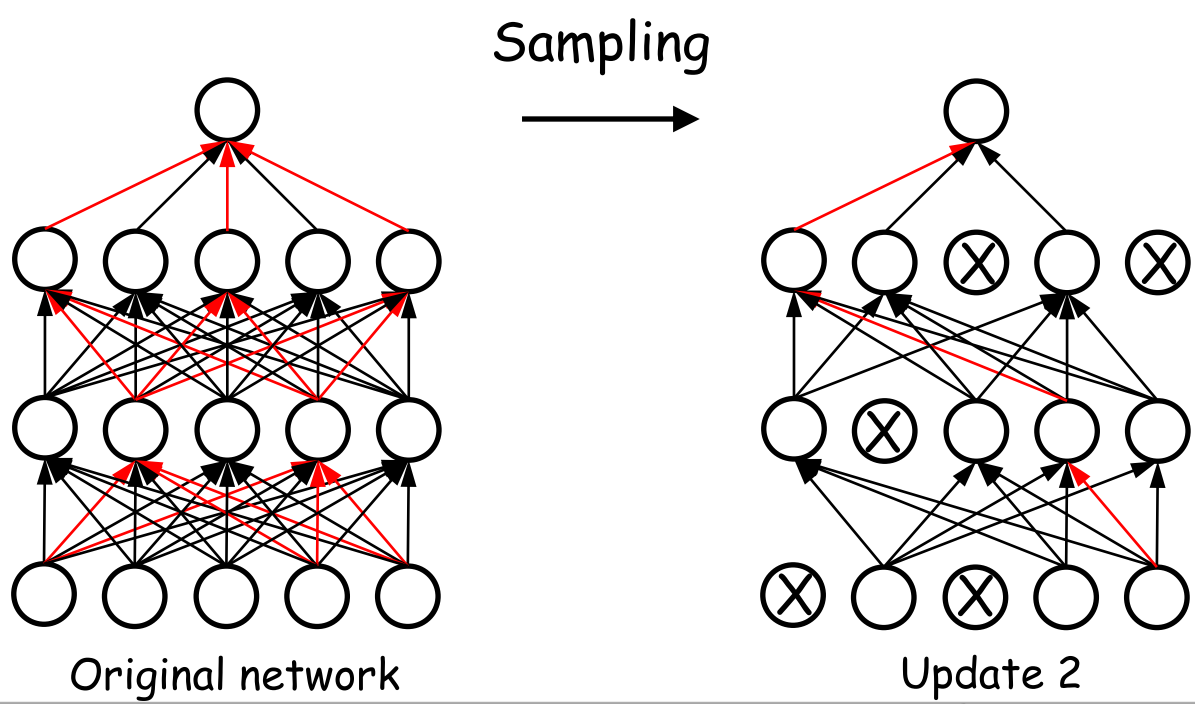
Dropout

- **How can we reduce the structural complexity without removing nodes?**
  - Hmm??



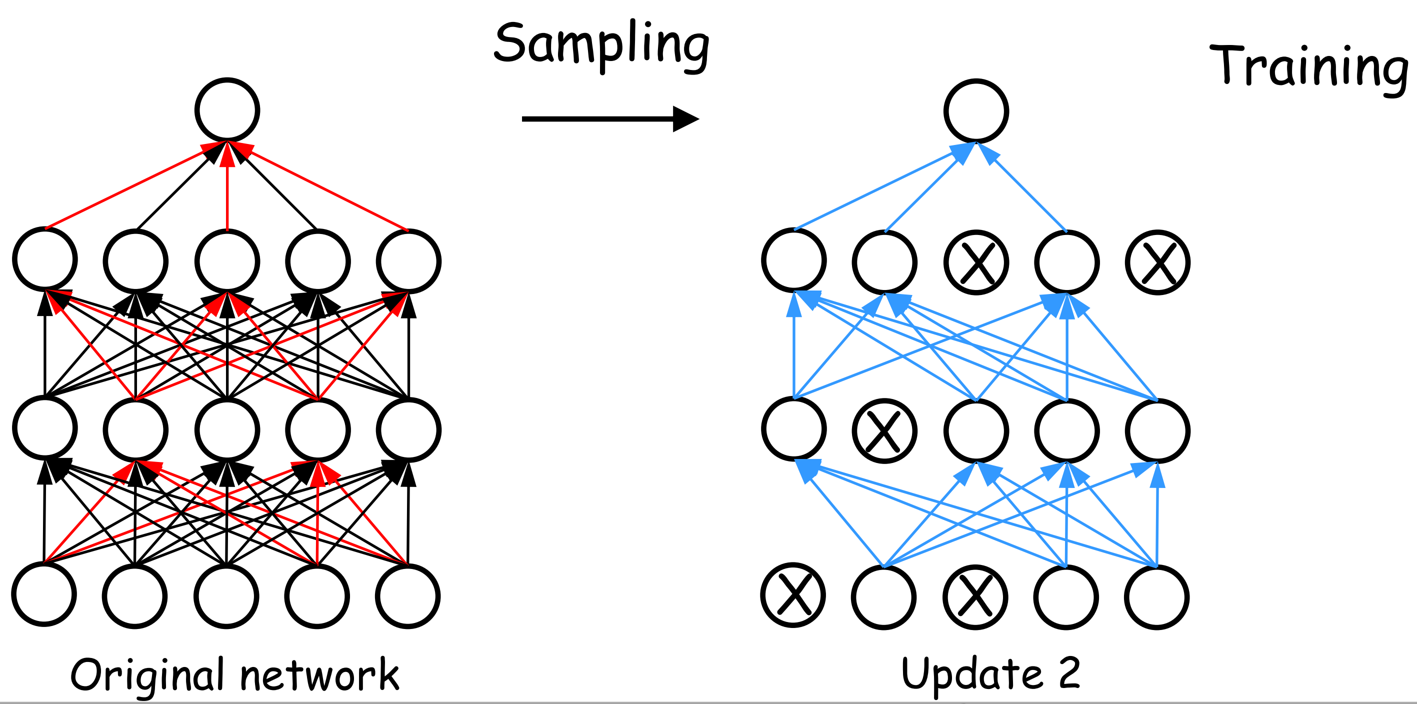
Dropout

- **How can we reduce the structural complexity without removing nodes?**
  - Hmm??



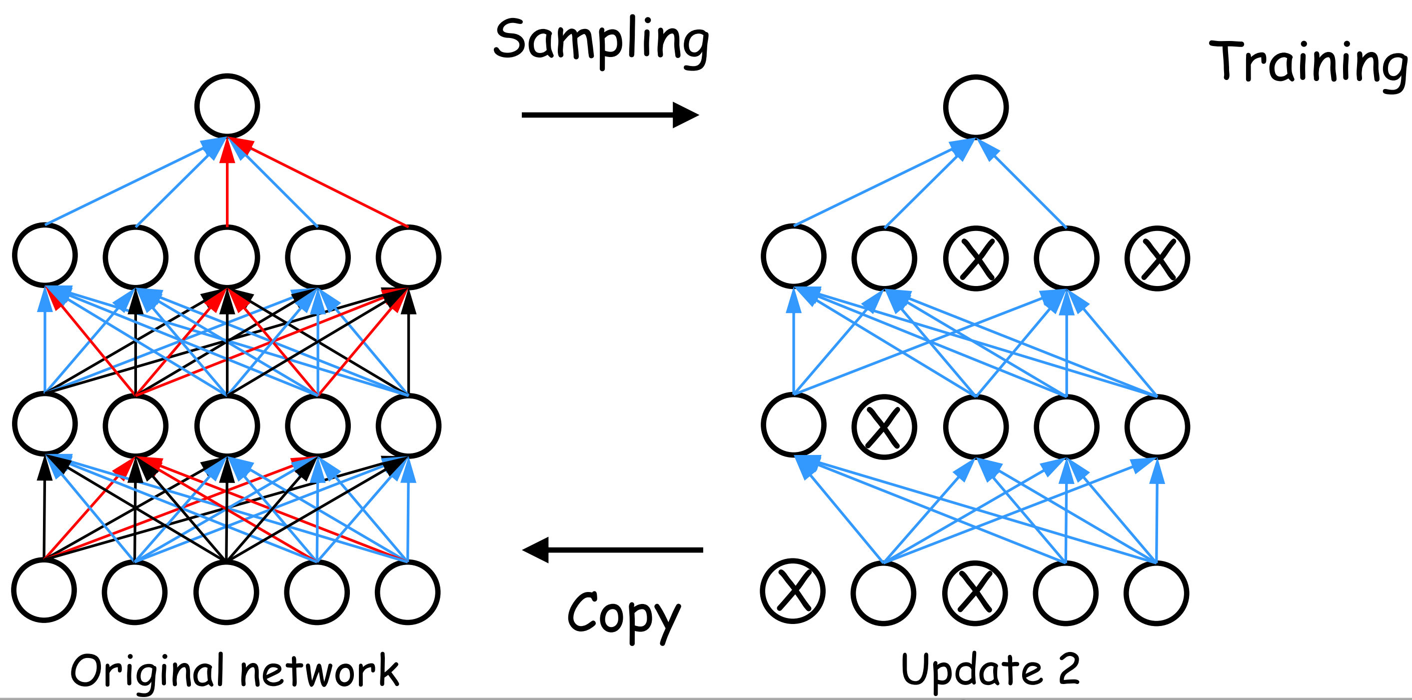
Dropout

- **How can we reduce the structural complexity without removing nodes?**
  - Hmm??



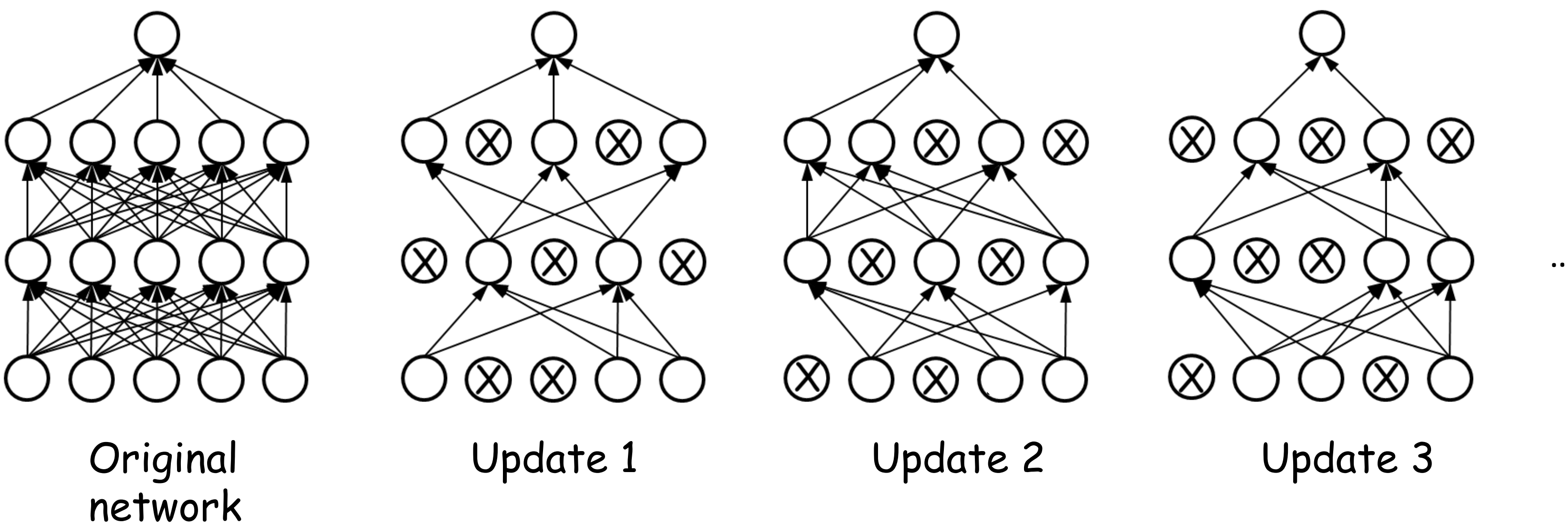
Dropout

- **How can we reduce the structural complexity without removing nodes?**
  - Hmm??



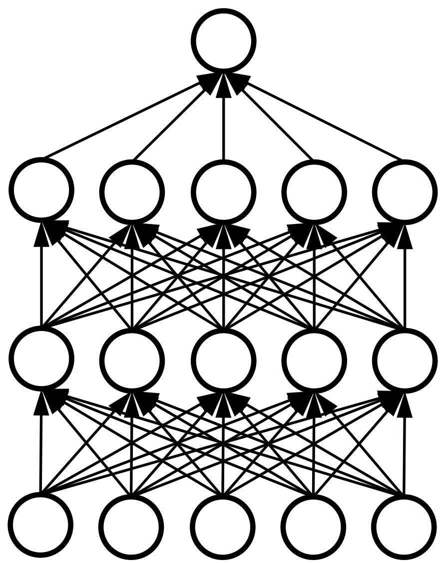
Dropout

- **Do this at every epoch**
  - Randomly choose nodes with a probability of *p*
    - Usually p = 0.5
  - Train the simplified neural network
    - At every epoch, we train different neural network which share connection weight each other



Dropout

- **Testing**
  - Use all the nodes without dropout

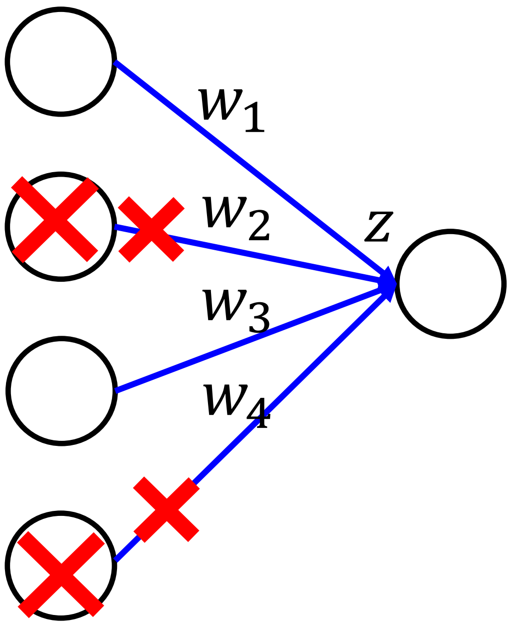


Dropout

- **Testing**

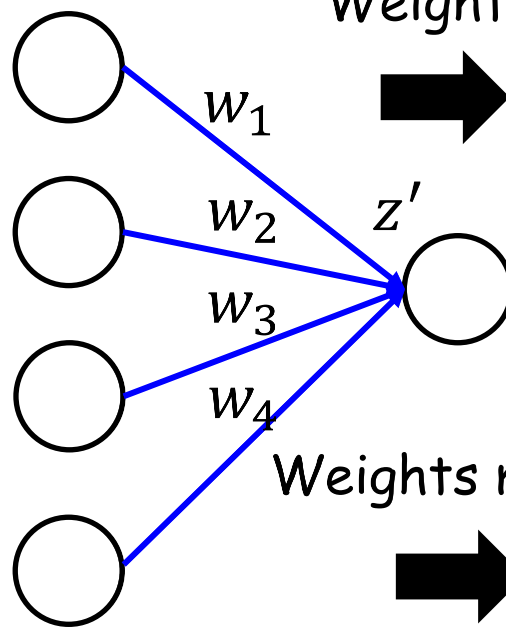
Training of Dropout

Assume dropout rate is 50%



Testing of Dropout

No dropout



Weights from training

$\Rightarrow z' \approx 2z$

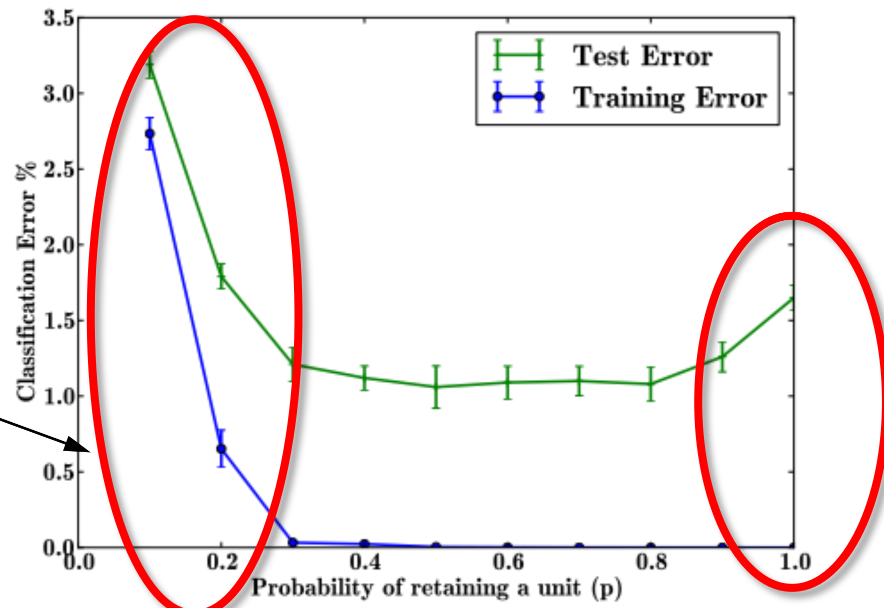
Weights multiply (1-p)%

$\Rightarrow z' \approx z$

Dropout

- **The effect of the dropout rate p:**
  - An architecture of 784-2048-2048-2048-10 is used on the MNIST dataset.
  - The dropout rate p is changed from small numbers (most units are dropped out) to 1.0 (no dropout).

Dropout with
high probability



No Dropout
->Overfitting

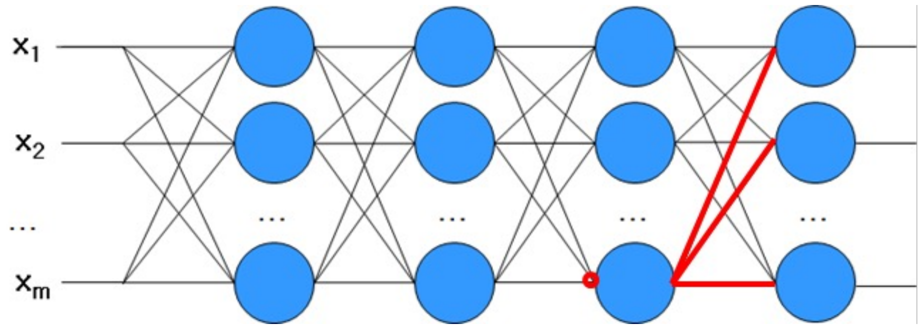
Dropout

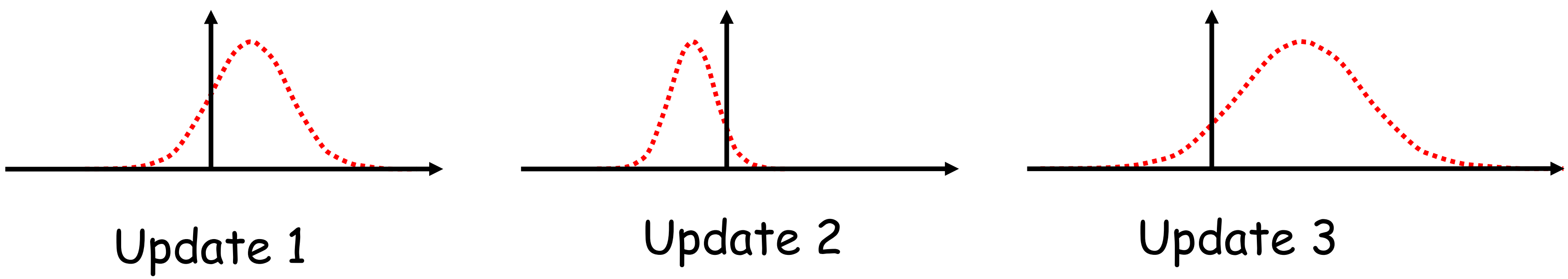
- **Summary**
  - Dropout is a very good and fast regularization method.
  - Dropout is a bit slow to train (2-3 times slower than without dropout).
  - If the amount of data is average-large – dropout excels. When data is big enough, dropout does not help much.
  - Dropout achieves better results than former used regularization methods (Weight Decay).

Batch Normalization

Batch Normalization

- **Distribution Shift**
  - Output distribution of the red node

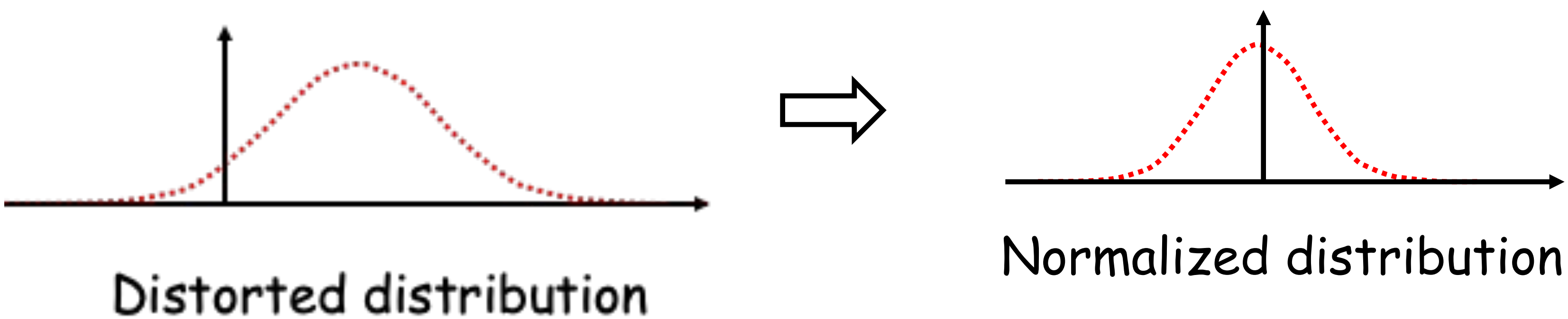




Batch Normalization

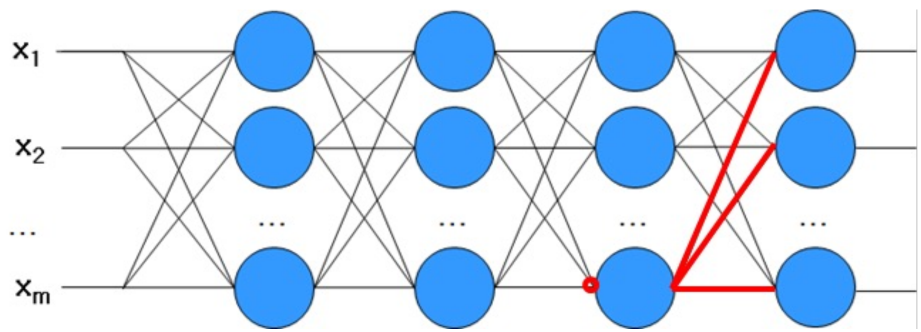
- **Distribution Shift**
  - It disturbs the learning process,
  - Learning is getting slow down

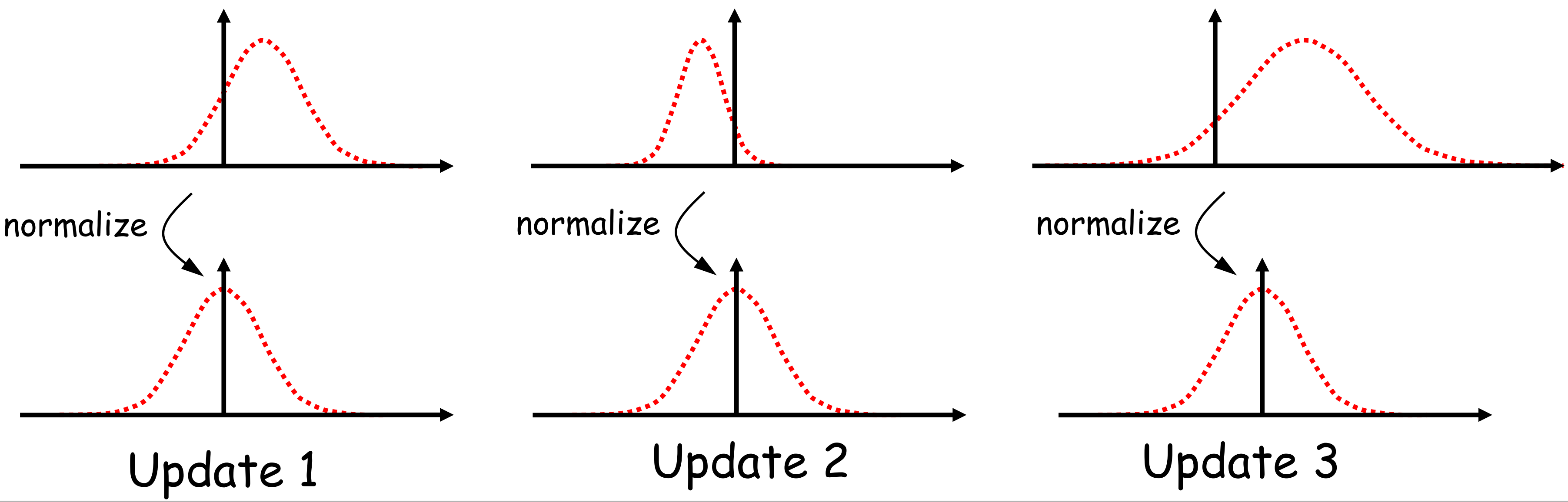
- **Why don't we normalize the distribution of inputs**



Batch Normalization

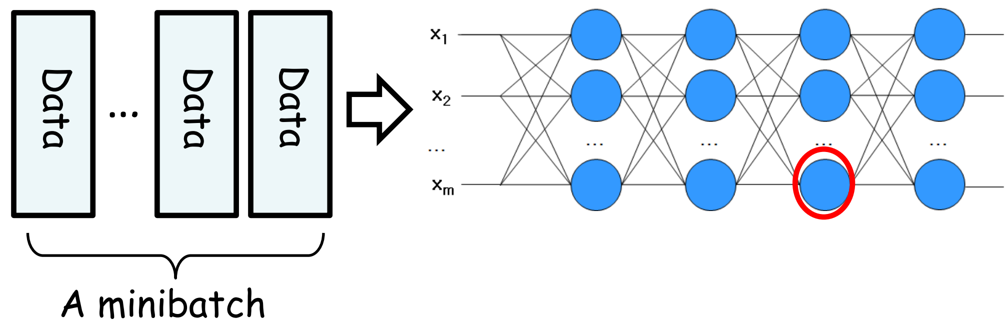
- **Normalization of outputs**

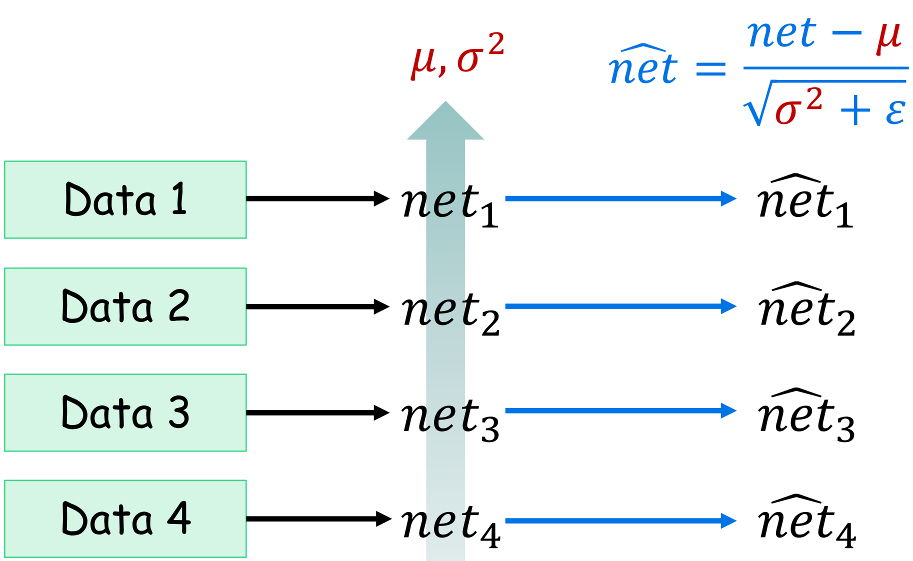


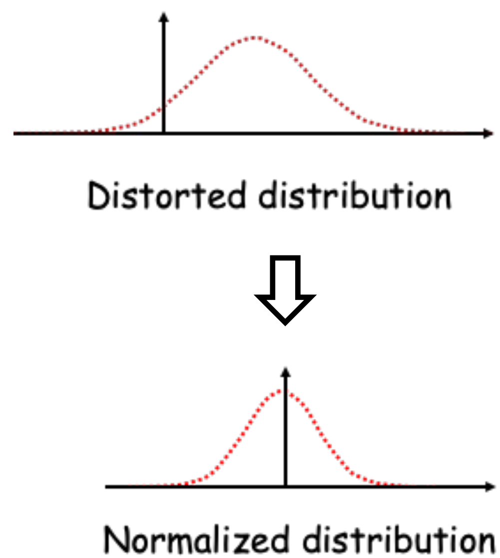


Batch Normalization

- **Input Normalization**

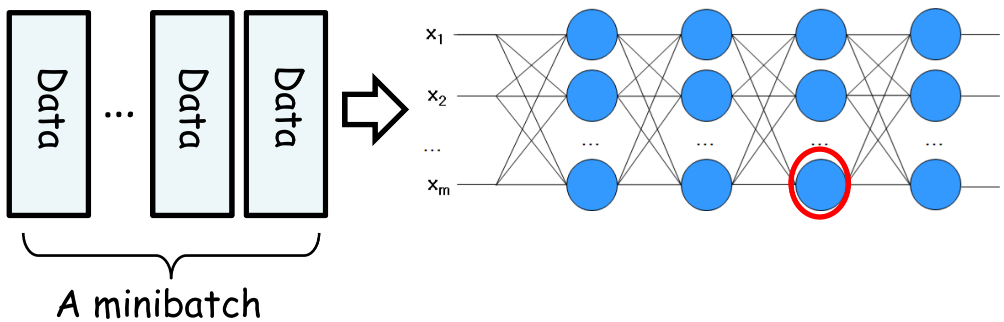






Batch Normalization

- **Input Normalization**



μ, σ^2

$$\widehat{net} = \frac{net - \mu}{\sqrt{\sigma^2 + \varepsilon}}$$

Data 1 → net_1 → $\widehat{net}_1$ → $f(\widehat{net}_1)$

Data 2 → net_2 → $\widehat{net}_2$ → $f(\widehat{net}_2)$

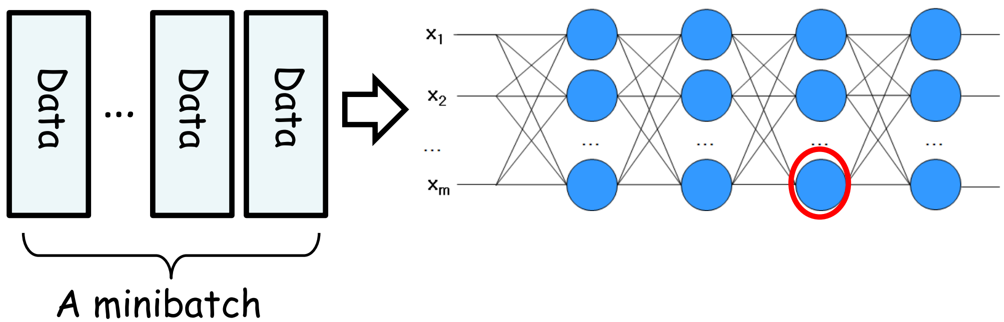
Data 3 → net_3 → $\widehat{net}_3$ → $f(\widehat{net}_3)$

Data 4 → net_4 → $\widehat{net}_4$ → $f(\widehat{net}_4)$

?

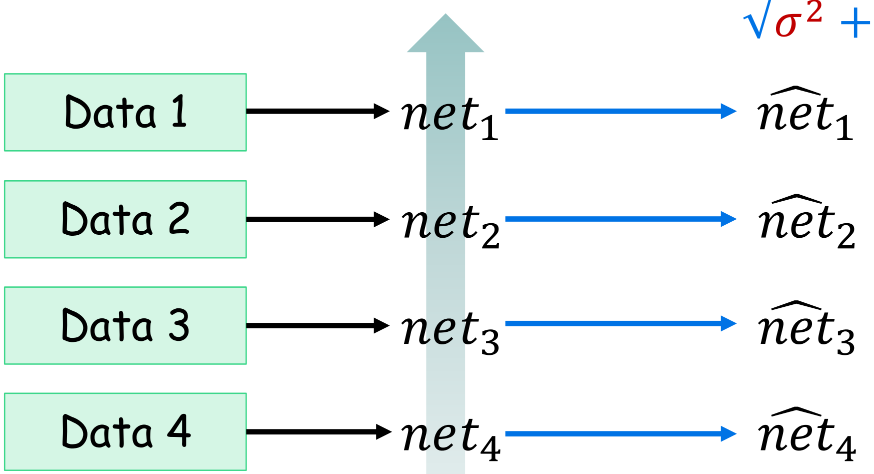
Batch Normalization

- **Input Normalization**



μ, σ^2

$$\widehat{net} = \frac{net - \mu}{\sqrt{\sigma^2 + \varepsilon}}$$



$$net = \mathbf{wx} + b$$

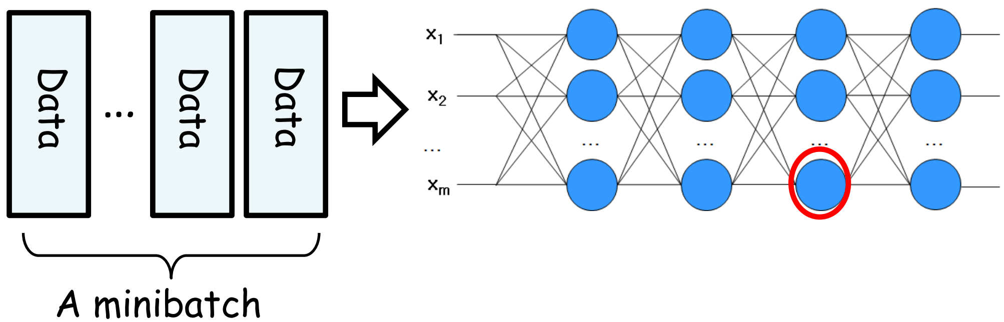
$$E(net) = E(\mathbf{wx}) + b$$

$$\widehat{net} = \frac{\mathbf{wx} + b - (E(\mathbf{wx}) + b)}{\sqrt{\sigma^2 + \varepsilon}}$$

$$\widehat{net} = \frac{\mathbf{wx} - E(\mathbf{wx})}{\sqrt{\sigma^2 + \varepsilon}}$$

Batch Normalization

▪ Input Normalization



μ, σ^2

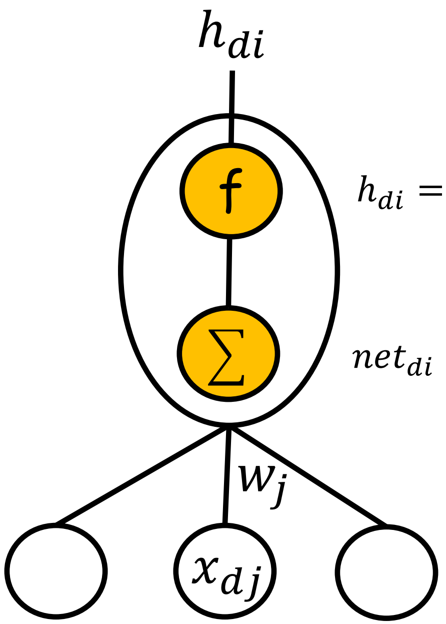
$$\widehat{net} = \frac{net - \mu}{\sqrt{\sigma^2 + \varepsilon}}$$

$$\widetilde{net} = \gamma \widehat{net} + \beta$$

Data 1	net_1	$\widehat{net}_1$	$\widetilde{net}_1$	$f(\widetilde{net}_1)$
Data 2	net_2	$\widehat{net}_2$	$\widetilde{net}_2$	$f(\widetilde{net}_2)$
Data 3	net_3	$\widehat{net}_3$	$\widetilde{net}_3$	$f(\widetilde{net}_3)$
Data 4	net_4	$\widehat{net}_4$	$\widetilde{net}_4$	$f(\widetilde{net}_4)$

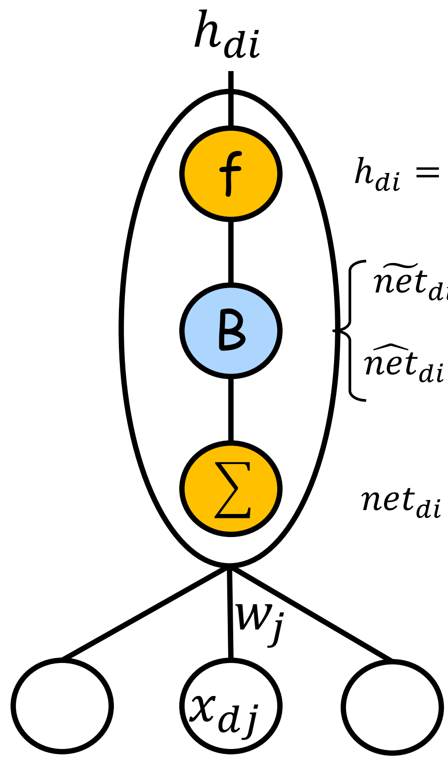
Batch Normalization

- **For a Single Node**



$$h_{di} = Activation(net_{di})$$

$$net_{di} = \sum_{j=1}^{J} w_j x_{dj} + b$$



$$h_{di} = Activation(\widetilde{net}_{di})$$

$$\begin{cases} \widetilde{net}_{di} = \gamma \widehat{net}_{di} + \beta \\ \widehat{net}_{di} = \dfrac{net_{di} - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \end{cases}$$

$$net_{di} = \sum_{j=1}^{J} w_j x_{dj} + b$$

$$\mu_B = \frac{1}{D_B} \sum_{d=1}^{D_B} net_{di}$$

$$\sigma_B^2 = \frac{1}{D_B} \sum_{d=1}^{D_B} (net_{di} - \mu)^2$$

Batch Normalization

- **Testing**
  - For Training, the mean and variance of each batch are used for normalization
  - For Testing, of which data the mean and variance will be used?
    - Estimated with those of batches in the training

$$\widehat{net}_{ti} = \frac{net_{di} - \mu}{\sqrt{\sigma^2 + \varepsilon}} \xleftarrow{\text{Transform}} net_{ti} = \sum_{j=1}^{J} w_j x_{tj} + b$$

Test sample

$$\mu = \frac{1}{B}\sum_{b=1}^{B} \mu_b$$

Unbiased estimator of mean of the population

$$\sigma^2 = \frac{1}{B}\sum_{b=1}^{B} \frac{m}{m-1}\sigma_b^2$$

Mean of unbiased estimator of variance of the population

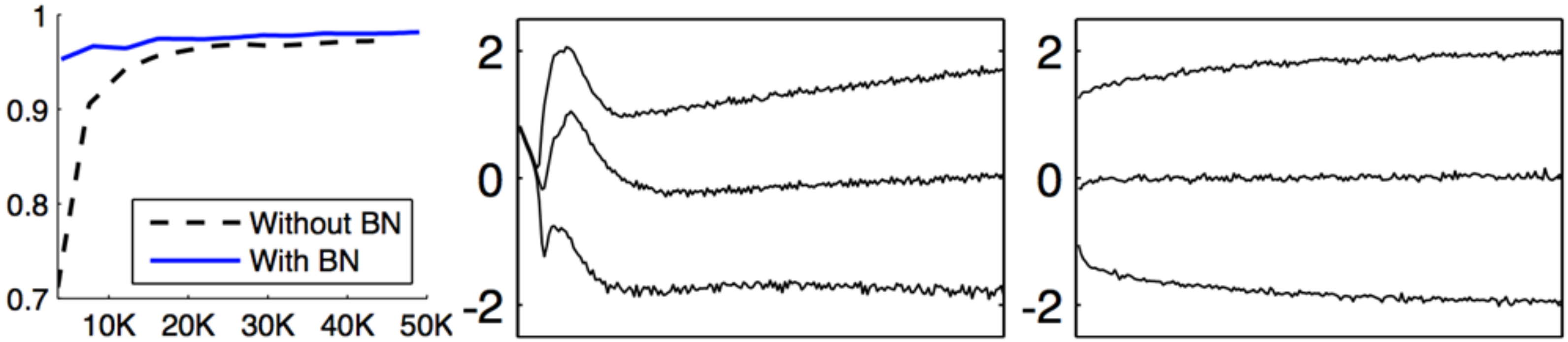
B: # of batches
m: # of samples in each batch

$$\widetilde{net}_{di} = \gamma \widehat{net}_{di} + \beta$$

$$h_{di} = Activation(\widetilde{net}_{di})$$

Batch Normalization

- **Performance with BN**



(a) *accuracy* (b) Without BN (c) With BN

input distributions

Batch Normalization

- **Advantage**
  - Reduces internal covariant shift.
  - Reduces the dependence of gradients on the scale of the connection weights.
  - Regularizes the model and reduces the need for regularization techniques.
    - It adds some stochastic noise to the activations as a result of using noisy estimates computed on the mini-batches. This has a regularization effect in some applications,

Batch Normalization

- **Disadvantage**
  - Expensive: Memory and time
    - Must keep interim results of all instances in a batch
    - Especially in CNN, usually an image is large
  - Hard to apply when the batch size is small
    - If batches are small, the means and variances cannot approximate the global ones.
  - Hard to apply to recurrent networks
    - It doesn't match to structure of recurrent networks
    - Hard to implement with recurrent networks