Deep Learning Notes

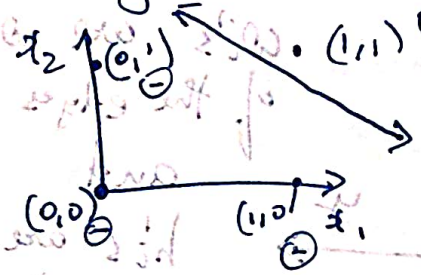
Harshil Pandey

2025

Contents

Linearly Separable Functions

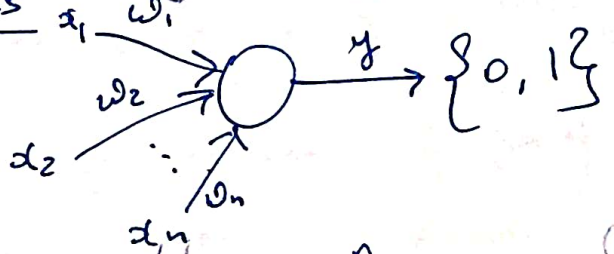
A linearly separable function is a boolean function that can be separated into different classes using a hyperplane in a high-dimensional space.



AND Function is linearly separable.

XOR is not a linearly separable function.

Perceptrons



$y = 1$ if $w_0 + \sum_{i=1}^{n} w_i x_i \geq 0$

$y = 0$ otherwise

a) w_0 is called the bias. if $w_0 \geq 0$, we support +ve class more.

=> We say that the perceptron fires if $w_0 + \sum_{i=1}^{n} x_i w_i \geq 0$ is true.

$\{True, False\}^n \rightarrow \{True, False\}$ can be implemented using a network of perceptrons.

Assume that True = +1 and False = -1

=> For 2 inputs the combinations can be $2^2 = 4$. Hence 2^4 boolean functions can be made for 2 inputs, each combination of input can take output of -1 or 1.

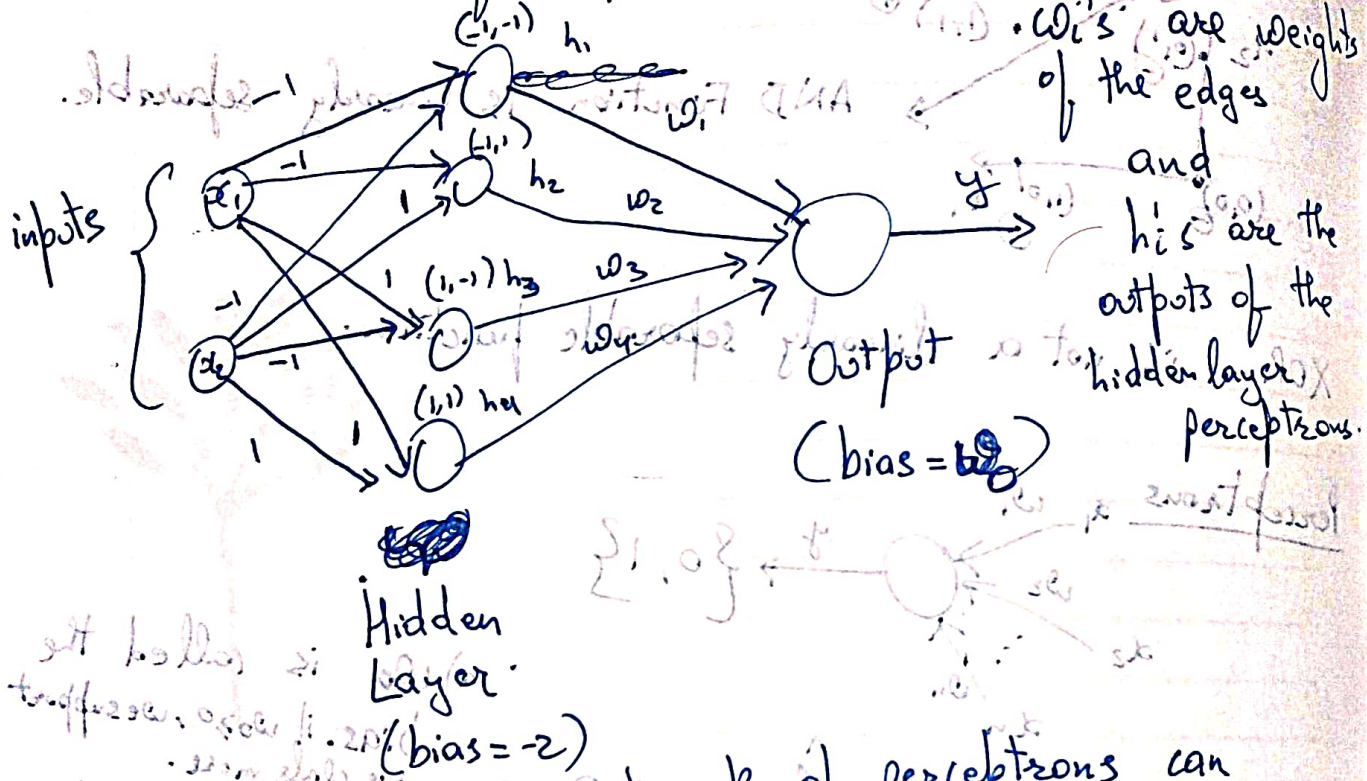
So for n inputs total possible boolean function: 2^{2^n}

=) Example for 2 inputs.

4 possible input combinations:-

$(-1,-1), (-1,1), (1,-1), (1,1)$

Let us make four perceptrons:-



w_i's are weights of the edges and h_i's are the outputs of the hidden layer perceptrons.

The claim is that this network of perceptrons can represent any boolean function of 2 inputs.

This is because if the input is $(-1,-1)$ only the top neuron will fire, as its output will be $(-1)(-1) + (-1)(-1) - 2$

$= 0 \geq 0$

None other neuron will fire because we have designed our network in such a way. Hence for each input only the corresponding neuron will fire. So whatever is the desired output for a particular input, that corresponding weight from the corresponding neuron to the output neuron can be adjusted. For example if the desired output from the input $(1,-1)$ is 1, then, only third neuron will

first, hence w_3 can be adjusted w.r.t w_0 such that $w_3 \geq w_0$. Since no other neuron fires, w_3 can be adjusted independently as it will be the only one contributing to $\sum_{i=1}^{4} w_i x_i$ w.r.t w_0 when input is $(1,-1)$, also w_3 will never contribute whenever input is other than $(1,-1)$ so it will not matter what w_3 is in those cases.

=> When there are 3 inputs, 8 possible input combinations are there. So we use 8 perceptrons in the hidden layer with bias = -3, each firing only to 1 possible input.

Representation power of a multilayer network of perceptrons

=> A multilayer network of perceptrons with a single hidden layer can be used to represent any boolean function precisely (no errors).

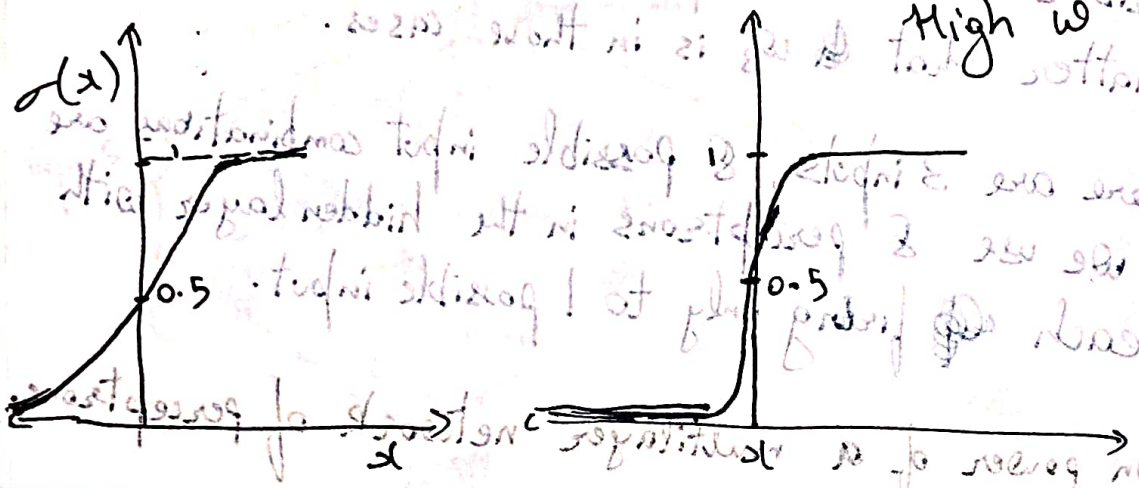
Representation power of a multilayer network of Sigmoid neurons

=> A multilayer network of neurons with a single hidden layer can be used to approximate any continuous function to any desired precision. There is a guarantee that for any function $f(x): \mathbb{R}^n \to \mathbb{R}^m$, we can always find a neural network (with 1 hidden layer containing enough neurons) whose output $g(x)$ satisfies $|g(x) - f(x)| < \epsilon$

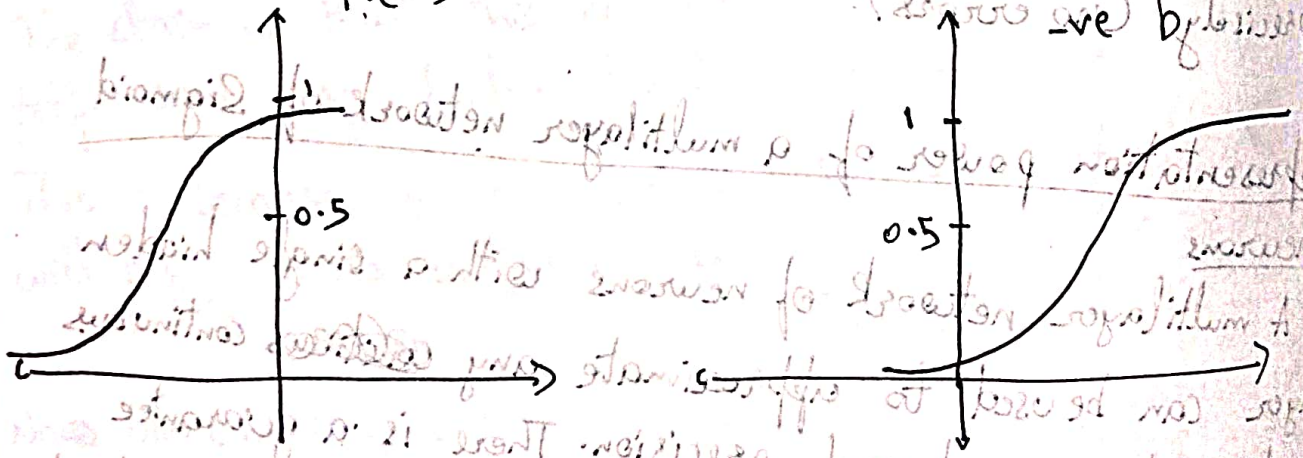
for any $\epsilon > 0$

⇒ Some facts about the shape of the sigmoid function (1-D):

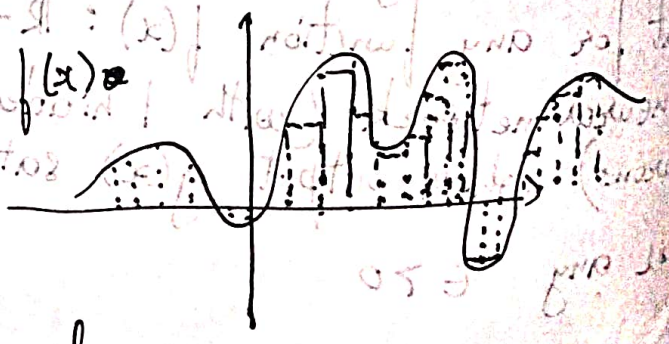
i) if $f(x) = \frac{1}{1+e^{-(wx+b)}}$, when $|w| \to \infty$ the function becomes closer and closer to the step function.



ii) b can varied to linearly translate the position of the curve.



Now Let there be a function $f(x)$ =

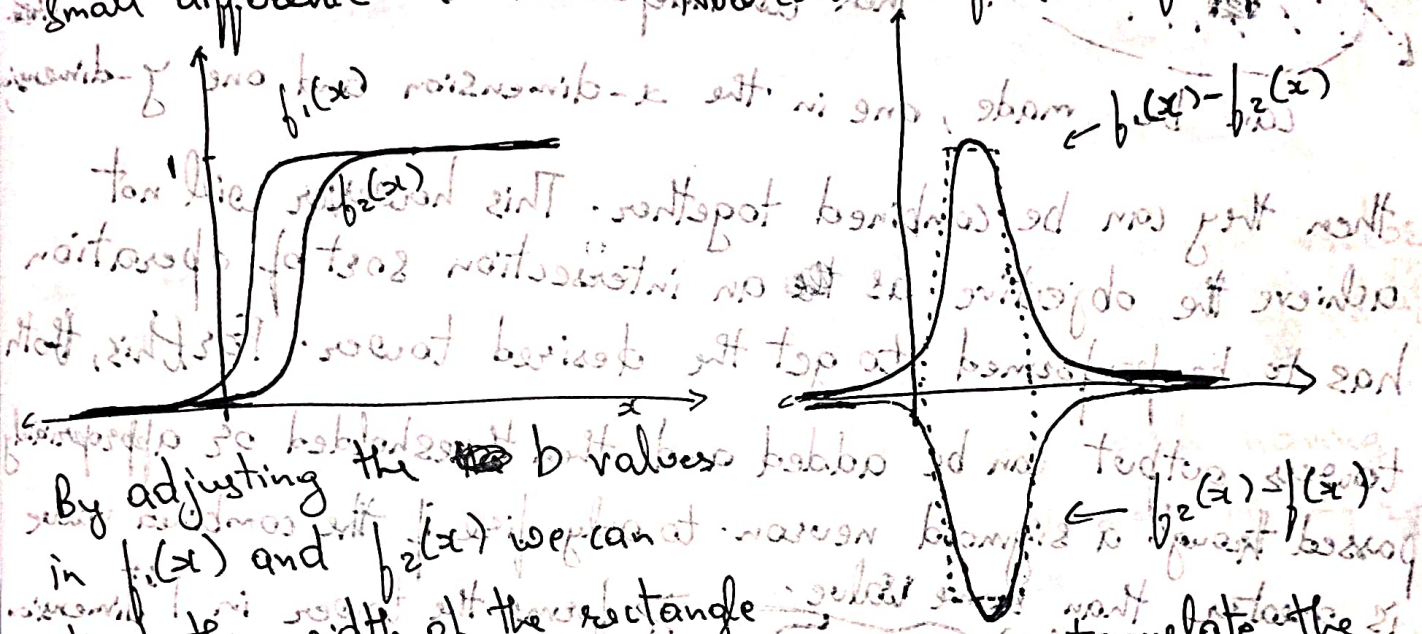


this can be approximated through equal width rectangles.

If we introduce more rectangles by reducing the width of the rectangles, then the approximation is more accurate.

The question is how to create these rectangle like structures using sigmoid neurons.

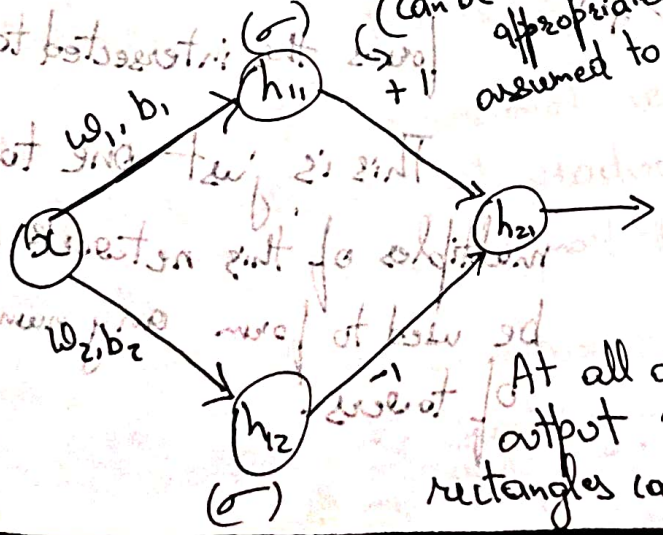
Consider two sigmoid functions with very high w and small difference in b. Now consider $f_1(x) - f_2(x)$



By adjusting the b values in $f_1(x)$ and $f_2(x)$ we can adjust the width of the rectangle like structure, we can also manipulate it to translate the structure linearly.

The height of the rectangle can be manipulated by scaling the output of the sigmoid appropriately.

Combining this into a neural network:—

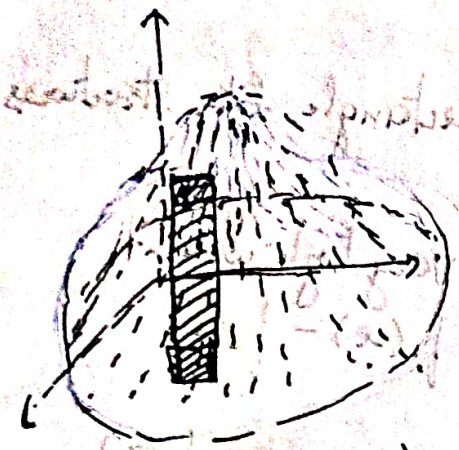


(can be scaled appropriately to achieve the correct height, assumed to be 1 for simplicity)

Notice that by adjusting w_i's and b_i's we can make a very specific rectangle at desired position and height. At all other x input, the network will output 0 and won't fire. Any number of rectangles can be made to approximate the function by using the multiples of this network.

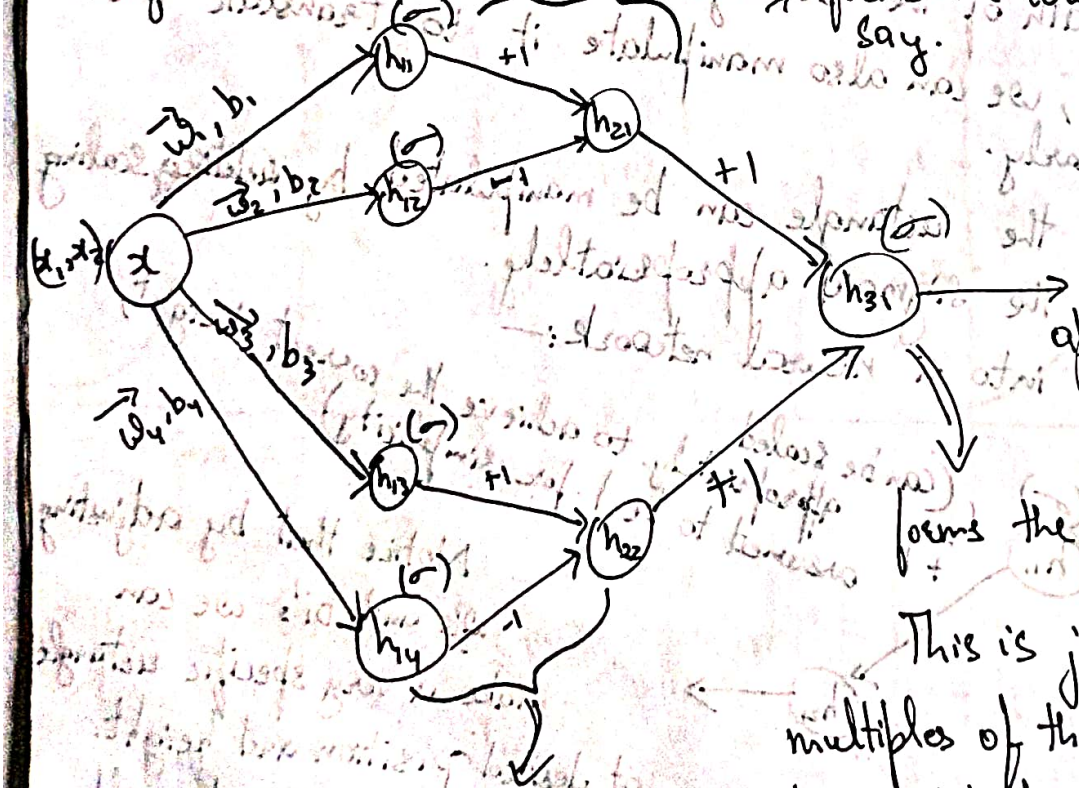
For 2-D case :- each dimension can have its own tower.

Suppose the tower has to be made between the area :- $\begin{bmatrix} (x_{min} - x_{max}) \\ (y_{min} - y_{max}) \end{bmatrix}$.



then two independent towers can be made, one in the x-dimension and one y-dimension

then they can be combined together. This however will not achieve the objective as an intersection sort of operation has to be performed to get the desired tower. For this, both towers output can be added and then thresholded or appropriately passed through a sigmoid neuron to only fire if the combined value is greater than some value. → forms the tower in 1st dimension say.



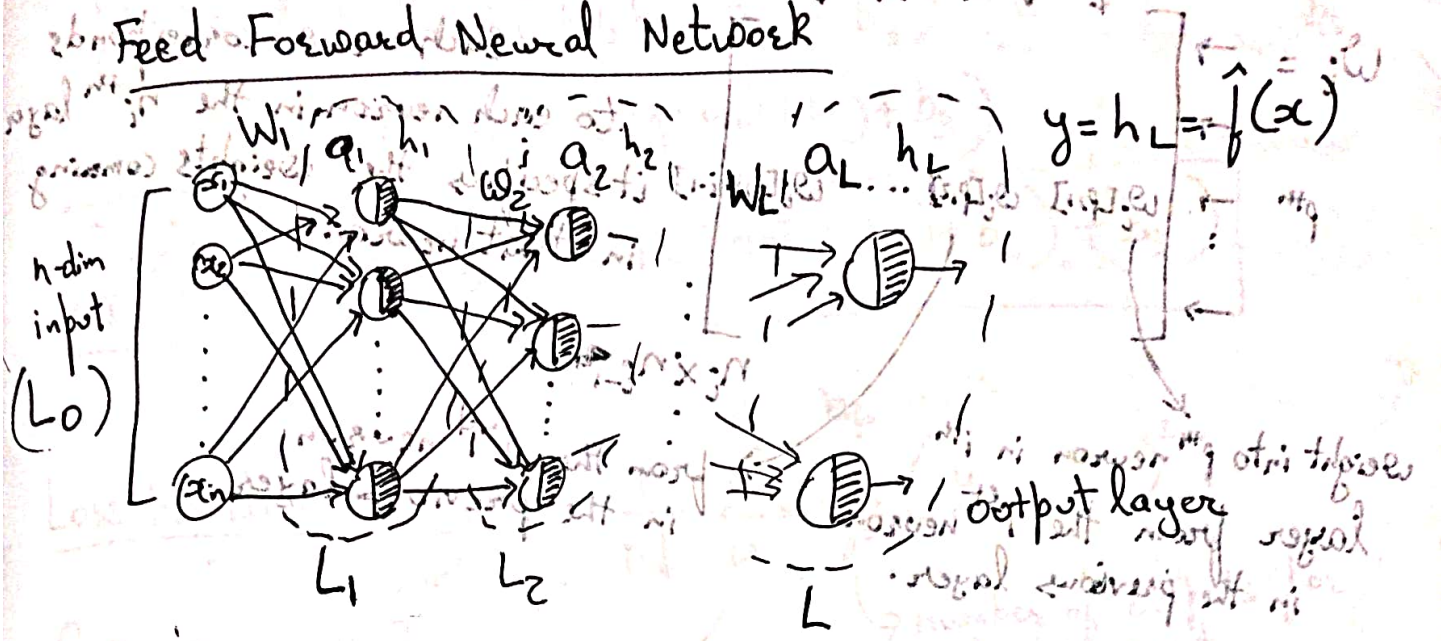
The hidden layer weights can be appropriately scaled.

forms the intersected tower.

forms the tower in 2nd dimension say.

This is just one tower, multiples of this network can be used to form any number of towers.

Feed Forward Neural Network



⇒ The input to the network is an n-dimensional vector. $x \in \mathbb{R}^n$. The input Layer is called the 0^{th} Layer of L_0.

⇒ The network contains L-1 hidden layers each having n_i neurons. $n_0 = n$ (input). i represents the layer index.

⇒ Each neuron in the hidden layer and output layer can be split into two parts: pre-activation (a_i) and activation (h_i).

⇒ a_i is an $\mathbb{R}^{n_i}$ dimensional vector such that the j^{th} component of a_i is the input given to the j^{th} neuron of the i^{th} layer L_i. It is equivalent to $w^T x + b$ for the j^{th} neuron.

⇒ h_i is an $\mathbb{R}^{n_i}$ dimensional vector such that the j^{th} component of h_i is the output produced by the j^{th} neuron of the i^{th} layer L_i. If neuron is sigmoid then it is equivalent to $\sigma(w^T x + b)$.

⇒ $w_i \in \mathbb{R}^{n_i \times n_{i-1}}$ is a matrix corresponding to layer i.

$w_i[p, q]$ is the weight given to the input from the q^{th} neuron in $L_{n_{i-1}}$ layer to p^{th} neuron in L_{n_i} layer.

$W_i = \begin{bmatrix} \downarrow & \downarrow & \cdots & \downarrow \\ w_i[1,1] & w_i[1,2] & \cdots & w_i[1,n_{i-1}] \\ \vdots & & & \\ w_i[p,1] & & & w_i[p,n_{i-1}] \end{bmatrix}$ ← p^{th} row

$n_i \times n_{i-1}$

Each row corresponds to each neuron in the n_i^{th} layer it specifies the weights coming in that neuron

$w_i[p,1]$: weight into p^{th} neuron in i^{th} layer from the 1st neuron in the previous layer.

$w_i[p,2]$: from the 2nd neuron in the previous layer.

⇒ The activation a_i of the i^{th} layer will be:

$$a_i = W_i \times h_{i-1} + b_i$$

$n_i \times n_{i-1}$ · $n_{i-1} \times 1$ → $n_i \times 1$

2) The activation of the i^{th} layer (h_i) is the element wise function of a_i, for example an element wise sigmoid:-

$h_i = \vec{\sigma}(a_i)$ where $\vec{\sigma} : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$

such that $\vec{\sigma}(a_i)_j = \sigma(a_{ij})$

3) The activation at the output Layer is given by:

$\hat{f}(x) = h_L = O(a_L)$ ... (also an element-wise function)

Writing it as a function of x :- (assume only 2 hidden Layers are there)

$\hat{f}(x) = O(a_3)$

$= O(W_3 h_2 + b_3)$

$= O(W_3 g(a_2) + b_3)$

$$f(x) = O(W_3 g(W_2 h_1 + b_2) + b_3)$$
$$= O(W_3 g(W_2 g(a_1) + b_2) + b_3)$$

$$\boxed{f(x) = O(W_3 g(W_2 g(W_1 x + b_1) + b_2) + b_3)}$$

Loss Function:- $L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n_L} (\hat{y}_{ij} - y_{ij})^2$

$\{W_1, W_2, \cdots, W_L\}$, $\{b_1, b_2, \cdots, b_L\}$ → θ

N → number of input data points

n_L → number of outputs for each data point.

This is suitable for regression type problems.

Output Function for regression problems:- We would like to allow all real values as outputs so just a linear transformation can work nicely.

$h_L = O(a_L) = W_0 a_L + b_0$, ($W_0$ and b_0 are not the same as W_L and b_L)

For Multi Class Classification:- The true labels are $\{0, 1\}$ where only a single 1 is there in the output vector for 1 datapoint all other are zero. Can be thought of as a PMF with all its mass on the correct class. For modelling such pmf's the more appropriate output activation is the softmax function.

$$\hat{y}_j = O(a_L)_j = \frac{e^{a_{Lj}}}{\sum_{k=1}^{n_L} e^{a_{Lk}}}$$ → the j^{th} component of an output vector for 1 data point. $\hat{y}_i$ will $\in (0, 1)$

In this case, the more appropriate loss function is the cross-entropy loss

$$L(\theta) = -\sum_{j=1}^{n_L} y_j \log(\hat{y}_j) \quad \text{(Loss for 1 data point)}$$

$$\Rightarrow L(\theta) = -\log(\hat{y}_t) \quad \text{where } y_t = 1 \text{ (true label)}$$

and all other $y_j = 0$

$\hat{y}_j$ represents the probability of that data point that it belongs to the j^{th} class.

Gradient Computations (Cross entropy loss and softmax output neurons)

Let for some data point, the correct class is t^{th} class. Then the output vector for that point will be $e_t = [0 \; 0 \; \ldots \; 1 \; \ldots \; 0]$ (t^{th} position)

The loss corresponding to this point is:

$$L(\theta) = -\log(\hat{y}_t)$$

$\Rightarrow$ Let's calculate $\nabla_{h_L} L(\theta)$, gradient with respect to the output layer

$$\nabla_{h_L} L(\theta) = \nabla_{\hat{y}} L(\theta)$$

Let us start with $\frac{\partial L(\theta)}{\partial \hat{y}_j}$; the j^{th} element of $\nabla_{\hat{y}} L(\theta)$.

$$\Rightarrow \frac{\partial L(\theta)}{\partial \hat{y}_j} = \frac{\partial(-\log(\hat{y}_t))}{\partial \hat{y}_j} = \mathbb{1}(t=j) \times \frac{-1}{\hat{y}_t}$$

$$= \begin{cases} -\frac{1}{\hat{y}_t} & \text{if } j = t \\ 0 & \text{if } j \neq t \end{cases}$$

So $\nabla_{\hat{y}} L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial \hat{y}_1} \\ \frac{\partial L(\theta)}{\partial \hat{y}_2} \\ \vdots \\ \frac{\partial L(\theta)}{\partial \hat{y}_{n_L}} \end{bmatrix} = \frac{-1}{\hat{y}_t} \begin{bmatrix} \mathbb{1}(1=t) \\ \mathbb{1}(2=t) \\ \vdots \\ \mathbb{1}(n_L=t) \end{bmatrix} = -\frac{1}{\hat{y}_t} \times e_t$

$\therefore \boxed{\nabla_{\hat{y}} L(\theta) = \frac{-1}{\hat{y}_t} \times e_t}$

⇒ Now Let's calculate the gradient based on the the pre-activation layer of the output. $\nabla_{a_L} L(\theta)$.

For that we know that :- $\hat{y}_j = \frac{e^{a_{Lj}}}{\sum_{k=1}^{n_L} e^{a_{Lk}}}$ } So for t max.

$L(\theta) = -\log(\hat{y}_t)$

$\frac{\partial L(\theta)}{\partial a_{Lj}} = \frac{\partial(-\log(\hat{y}_t))}{\partial a_{Lj}} = \frac{\partial(-\log(\hat{y}_t))}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial a_{Lj}}$

$\Downarrow$

$\frac{\partial L(\theta)}{\partial \hat{y}_t} \times \frac{\partial \hat{y}_t}{\partial a_{Lj}} = -\frac{1}{\hat{y}_t} \times \frac{\partial \left(\frac{e^{a_{Lt}}}{\sum_{k=1}^{n_L} e^{a_{Lk}}} \right)}{\partial a_{Lj}}$

$= -\frac{1}{\hat{y}_t} \times \frac{\sum_{k=1}^{n_L} e^{a_{Lk}} \times \frac{\partial(e^{a_{Lt}})}{\partial a_{Lj}} - e^{a_{Lt}} \frac{\partial(\sum_{k=1}^{n_L} e^{a_{Lk}})}{\partial a_{Lj}}}{\left(\sum_{k=1}^{n_L} e^{a_{Lk}} \right)^2}$

$$= -\frac{1}{\hat{y}_t} \times \left(\frac{\sum_{k=1}^{n_L} e^{a_{Lk}} \; e^{a_{Lt}} \times \mathbb{1}(t=j) \; - \; e^{a_{Lt}} \; e^{a_{Lj}}}{\left(\sum_{k=1}^{n_L} e^{a_{Lk}}\right)^2} \right)$$

$$= -\frac{1}{\hat{y}_t} \times \left(\frac{\mathbb{1}(t=j) \, e^{a_{Lt}}}{\sum_{k=1}^{n_L} e^{a_{Lk}}} - \frac{e^{a_{Lt}} \; e^{a_{Lj}}}{\sum_{k=1}^{n_L} e^{a_{Lk}} \times \sum_{k=1}^{n_L} e^{a_{Lk}}} \right)$$

$$= -\frac{1}{\hat{y}_t} \left(\mathbb{1}(t=j) \, \hat{y}_t - \hat{y}_t \, \hat{y}_j \right)$$

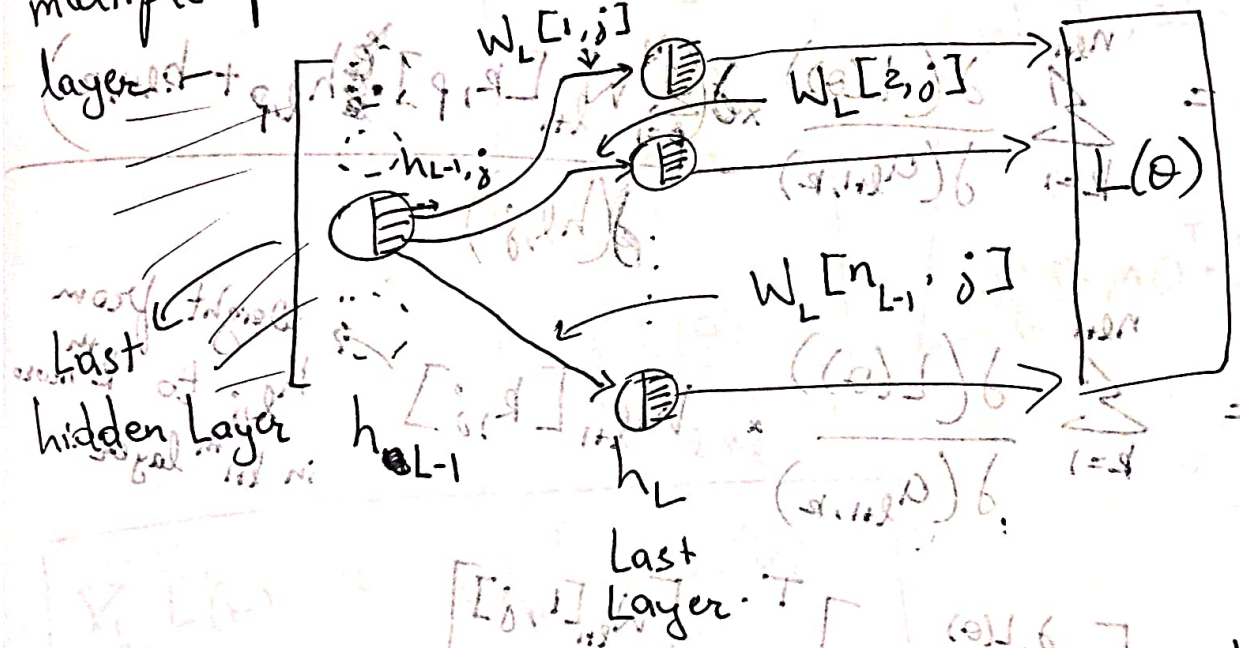
$$= -\left(\mathbb{1}(t=j) - \hat{y}_j \right)$$

$$\Rightarrow \frac{\partial L(\theta)}{\partial a_{Lj}} = -\left(\mathbb{1}(t=j) - \hat{y}_j \right)$$

Therefore: $\nabla_{a_L} L(\theta) = -\begin{bmatrix} \mathbb{1}(t=1) - \hat{y}_1 \\ \mathbb{1}(t=2) - \hat{y}_2 \\ \vdots \\ \mathbb{1}(t=n_L) - \hat{y}_{n_L} \end{bmatrix}$

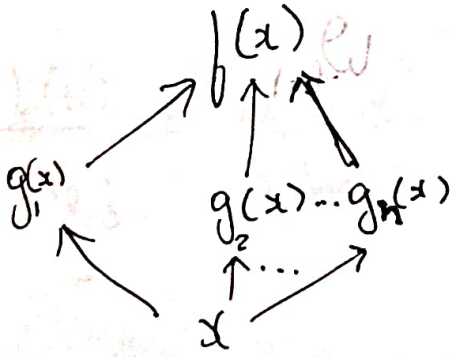
$$\boxed{\nabla_{a_L} L(\theta) = -(e_t - \hat{y})}$$

We now have the formulas for gradients of the output layer. This is straightforward as the output layer directly goes into the Loss function. This is not true however, for any other layer as the output of hidden layer affects the loss function through multiple paths. Let us take a neuron in the last hidden layer.



We can see that output of $h_{L-1,j}$ j^{th} neuron of the last hidden layer affects the path through L paths.

⇒ Hence we use chain rule for multiple path to calculate $\frac{\partial L(\theta)}{\partial h_{lj}}$ l is for the layer number and j is for the neuron number.



Let $f(x)$ be a function of x such that it can be separated as $g_1(x), g_2(x)$ and ... $g_n(x)$. Then

$$\frac{\partial f(x)}{\partial x} = \sum_{i=1}^{n} \frac{\partial f(x)}{\partial g_i(x)} \times \frac{\partial g_i(x)}{\partial x}$$

In our case $f(x) = L(\theta)$, $g_i(x) = a_{l+1,k}$, $x = h_{l,j}$

k^{th} pre-activation of the next layer.

$$\frac{\partial L(\theta)}{\partial h_{l,j}} = \sum_{k=1}^{n_{l+1}} \frac{\partial (L(\theta))}{\partial (a_{l+1,k})} \frac{\partial (a_{l+1,k})}{\partial (h_{l,j})}$$

$$= \sum_{k=1}^{n_{l+1}} \frac{\partial (L(\theta))}{\partial (a_{l+1,k})} \frac{\partial \left(\sum_{p=1}^{n_l} W_{l+1}[k,p] \cdot h_{l,p} + b_{l+1,k} \right)}{\partial (h_{l,j})}$$

$$= \sum_{k=1}^{n_{l+1}} \frac{\partial (L(\theta))}{\partial (a_{l+1,k})} \times W_{l+1}[k,j]$$

$W_{l+1}[k,j]$ → weight from $h_{l,j}$ to k^{th} neuron in $(l+1)^{th}$ layer

$$= \begin{bmatrix} \frac{\partial L(\theta)}{\partial a_{l+1,1}} \\ \frac{\partial L(\theta)}{\partial a_{l+1,2}} \\ \vdots \\ \frac{\partial L(\theta)}{\partial a_{l+1,n_{l+1}}} \end{bmatrix}^T \cdot \begin{bmatrix} W_{l+1}[1,j] \\ W_{l+1}[2,j] \\ \vdots \\ W_{l+1}[n_{l+1},j] \end{bmatrix}$$

$\frac{\partial L(\theta)}{\partial a_{l+1,n_{l+1}}}$ → $\nabla_{a_{l+1}} L(\theta)$

$\begin{bmatrix} W_{l+1}[1,j] \\ \vdots \\ W_{l+1}[n_{l+1},j] \end{bmatrix}$ → j^{th} column of W_{l+1}

$$\Rightarrow \frac{\partial L(\theta)}{\partial h_{l,j}} = W_{l+1}[\bullet, j]^T \cdot \nabla_{a_{l+1}} L(\theta)$$

j^{th} column

$$\nabla_{h_l} L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial h_{l,1}} \\ \frac{\partial L(\theta)}{\partial h_{l,2}} \\ \vdots \\ \frac{\partial L(\theta)}{\partial h_{l,n_l}} \end{bmatrix} = \begin{bmatrix} W_{l+1}[\bullet, 1]^T \nabla_{a_{l+1}} L(\theta) \\ W_{l+1}[\bullet, 2]^T \cdot \nabla_{a_{l+1}} L(\theta) \\ \vdots \\ W_{l+1}[\bullet, n_l]^T \cdot \nabla_{a_{l+1}} L(\theta) \end{bmatrix}$$

$$\boxed{\nabla_{h_l} L(\theta) = W_{l+1}^T \cdot \nabla_{a_{l+1}} L(\theta)}$$

$\Rightarrow$ Now let us look at the $\nabla_{a_l} L(\theta)$. We know that

$h_l = g(a_l)$

$h_{l,j} = g(a_{l,j})$ (Element wise function)

g is the activation function

$$\frac{\partial L(\theta)}{\partial a_{l,j}} = \frac{\partial L(\theta)}{\partial h_{l,j}} \frac{\partial (h_{l,j})}{\partial a_{l,j}} = \frac{\partial (L(\theta))}{\partial (h_{l,j})} g'(a_{l,j})$$

$$\nabla_{a_\ell} L(\theta) = \begin{bmatrix} \frac{\partial (L(\theta))}{\partial (a_{\ell,1})} \\ \frac{\partial (L(\theta))}{\partial (a_{\ell,2})} \\ \vdots \\ \frac{\partial (L(\theta))}{\partial (a_{\ell,n_\ell})} \end{bmatrix} = \begin{bmatrix} \frac{\partial L(\theta)}{\partial h_{\ell,1}} \times g'(a_{\ell,1}) \\ \frac{\partial L(\theta)}{\partial h_{\ell,2}} \times g'(a_{\ell,2}) \\ \vdots \\ \frac{\partial L(\theta)}{\partial h_{\ell,n_\ell}} \times g'(a_{\ell,n_\ell}) \end{bmatrix}$$

$$\nabla_{a_\ell} L(\theta) = \nabla_{h_\ell} L(\theta) \odot [\dots, g'(a_{\ell,j}) \dots]$$

Element wise multiplication of $\nabla_{h_\ell} L(\theta)$ and $g'(a_\ell)$ also called as Hadamard Product

Chronology of Gradient Computation

Calculate :- 1) $\nabla_{h_L} L(\theta)$ 2) $\nabla_{a_L} L(\theta)$

Would required

3) $\nabla_{h_{L-1}} L(\theta)$ ← 4) $\nabla_{a_{L-1}} L(\theta)$

5) $\nabla_{h_{L-2}} L(\theta)$ ← 6) $\nabla_{a_{L-2}} L(\theta)$

$\nabla_{h_1} L(\theta)$ ← $\nabla_{a_1} L(\theta)$

⇒ Computing gradients for the parameters. $\nabla_{W_k} L(\theta)$ $n_k \times n_{k-1}$ and $\nabla_{b_k} L(\theta)$ $n_k \times 1$

We know that :-

$$a_l = W_l h_{l-1} + b_l$$

$$\frac{\partial L(\theta)}{\partial W_l[j,k]}$$

$W_l[j,k]$

$$\nabla_{W_l} L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial W_l[1,1]} & \frac{\partial L(\theta)}{\partial W_l[1,2]} & \cdots & \frac{\partial L(\theta)}{\partial W_l[1,n_{l-1}]} \\ \frac{\partial L(\theta)}{\partial W_l[2,1]} & \frac{\partial L(\theta)}{\partial W_l[2,2]} & \cdots & \frac{\partial L(\theta)}{\partial W_l[2,n_{l-1}]} \\ \vdots & & & \\ \frac{\partial L(\theta)}{\partial W_l[n_l,1]} & \frac{\partial L(\theta)}{\partial W_l[n_l,2]} & \cdots & \frac{\partial L(\theta)}{\partial W_l[n_l,n_{l-1}]} \end{bmatrix}$$

Note that $W_l[j,k]$ only comes in the computation of $a_{l,j}$. Hence $W_l[j,k]$ affects $L(\theta)$ only through 1 path.

Hence:- $$\frac{\partial L(\theta)}{\partial W_l[j,k]} = \frac{\partial L(\theta)}{\partial a_{l,j}} \times \frac{\partial a_{l,j}}{\partial W_l[j,k]}$$

$$= \frac{\partial L(\theta)}{\partial a_{l,j}} \times \frac{\partial \left(\sum_{p=1}^{n_{l-1}} W_l[j,p] \cdot h_{l-1,p} + b_{l,j} \right)}{\partial W_l[j,k]}$$

Weight to j^{th} neuron from k^{th} neuron

$$\frac{\partial L(\theta)}{\partial W_l[j,k]} = \frac{\partial L(\theta)}{\partial a_{l,j}} \times h_{l-1,k}$$

→ derivative w.r.t. j^{th} pre-act

→ output from k^{th} neuron.

$$\begin{bmatrix} \frac{\partial L(\theta)}{\partial a_{l,1}} h_{l-1,1} & \frac{\partial L(\theta)}{\partial a_{l,1}} h_{l-1,2} & \cdots & \frac{\partial L(\theta)}{\partial a_{l,1}} h_{l-1,n_{l-1}} \\ \frac{\partial L(\theta)}{\partial a_{l,2}} h_{l-1,1} & \frac{\partial L(\theta)}{\partial a_{l,2}} h_{l-1,2} & & \frac{\partial L(\theta)}{\partial a_{l,2}} h_{l-1,n_{l-1}} \\ \vdots & & & \\ \frac{\partial L(\theta)}{\partial a_{l,n_l}} h_{l-1,1} & \frac{\partial L(\theta)}{\partial a_{l,n_l}} h_{l-1,2} & & \frac{\partial L(\theta)}{\partial a_{l,n_l}} h_{l-1,n_{l-1}} \end{bmatrix}$$

⇒ $\nabla_{W_l} L(\theta)$ =

*

*

$$\Rightarrow \nabla_{W_\ell} L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial a_{\ell,1}} \\ \frac{\partial L(\theta)}{\partial a_{\ell,2}} \\ \vdots \\ \frac{\partial L(\theta)}{\partial a_{\ell,n_\ell}} \end{bmatrix} \begin{bmatrix} h_{\ell-1,1} & h_{\ell-1,2} & \cdots & h_{\ell-1,n_{\ell-1}} \end{bmatrix}$$

$$\boxed{\nabla_{W_\ell} L(\theta) = \nabla_{a_\ell} L(\theta) \cdot h_{\ell-1}^T}$$

Similar argument as $\frac{\partial L(\theta)}{\partial W_\ell[j,k]}$

For $\nabla_{b_\ell} L(\theta)$ notice that $\frac{\partial L(\theta)}{\partial b_{\ell,i}} = \frac{\partial L(\theta)}{\partial a_{\ell,i}} \times \frac{\partial a_{\ell,i}}{\partial b_{\ell,i}}$

$$= \frac{\partial L(\theta)}{\partial a_{\ell,i}} \times 1$$

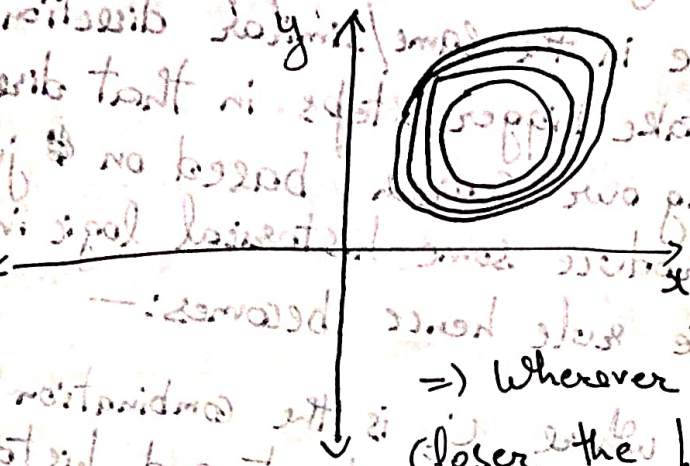
So $\boxed{\nabla_{b_\ell} L(\theta) = \nabla_{a_\ell} L(\theta)}$

$\Rightarrow \sigma(x) = \frac{1}{1+e^{-x}}$ $\quad \sigma'(x) = \sigma(x)(1-\sigma(x))$

$\Rightarrow \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ $\quad \tanh'(x) = 1 - (\tanh(x))^2$

Contours

Suppose there is a function from $\mathbb{R}^2 \to \mathbb{R}$, $z = f(x,y)$, then the contour map may look like:-



Each contour rings states the (x,y) values where the $f(x,y)$ is the same.

=> Wherever the contour rings are closer the function has a steep curvature

=> Wherever the contour rings are further apart, the function has a gentle slope/curvature.

The logic is that in areas of steep curvature, a small change in x and y values may bring a larger change in $f(x,y)$. Similarly in areas of gentle slopes a large change is required to bring the same amount of change in $f(x,y)$.

Δz change is achieved in small changes in x and y, steep curvature, contour rings are closer.

Δz change is achieved in larger changes in x and y, gentle slope, contour rings are further apart.

=> When gradient descent is applied, it makes tiny progress in areas where contour rings are far apart as the slope is small. It makes larger steps in areas where contour rings are closer as the slope is larger.

Momentum Based Gradient Descent

The small progress made in the flat areas causes the gradient descent to converge slowly. But if we are repeatedly asked to move in the same/similar directions, then it makes sense to take bigger steps in that direction. So instead of just guiding our search based on just the gradient, we introduce some historical logic into the search. So the update rule hence becomes:—

$$w_{t+1} = w_t - \eta v_t$$

where v_t is the combination of the current gradient and historical directions.

where,

$$v_t = \nabla w_t + \beta v_{t-1}$$

we start with w_0 and $v_0 = \nabla w_0$, $0 \le \beta < 1$.

Suppose we want w_3, then:— $w_3 = w_2 - \eta v_2$

$$v_2 = \nabla w_2 + \beta v_1$$

$$v_1 = \nabla w_1 + \beta v_0 = \nabla w_1 + \beta \nabla w_0$$

$$\therefore w_3 = w_2 - \eta(\nabla w_2 + \beta(\nabla w_1 + \beta \nabla w_0))$$

$$= w_2 - \eta(\nabla w_2 + \beta \nabla w_1 + \beta^2 \nabla w_0)$$

$$= w_2 - \eta \sum_{i=0}^{2} \beta^{2-i} \nabla w_i$$

$$\therefore w_{t+1} = w_t - \eta \underbrace{\sum_{i=0}^{t} \beta^{t-i} \nabla w_i}_{v_t}$$

you give the most importance to ∇w_2 which is the current gradient then ∇w_1, and then the least importance to ∇w_0.

So if we are making small steps in some direction, after some steps the "momentum" (v_t) will grow and we will start making bigger steps.

The problem with this strategy is that sometimes the momentum is too large and we overshoot unneccessarily in some directions. But it certainly converges faster than normal G.D. in most situations.

To deal with the overshooting problem of Momentum based gradient descent we look at the Nesterov Accelerated Gradient Descent (NAG). Instead of directly combining the historical direction v_{t-1} with the current gradient to calculate v_t we check what would happen if ∇w_t was not there. Then v_t will be βv_{t-1} :- $v_t = \beta v_{t-1}$

So $w'_{t+1} = w_t - \eta v_t = w_t - \eta \beta v_{t-1}$

if we ignore the step size η, then w'_{t+1} will be $w_t - \beta v_{t-1}$

Let us calculate the gradient w.r.t this w'_{t+1}

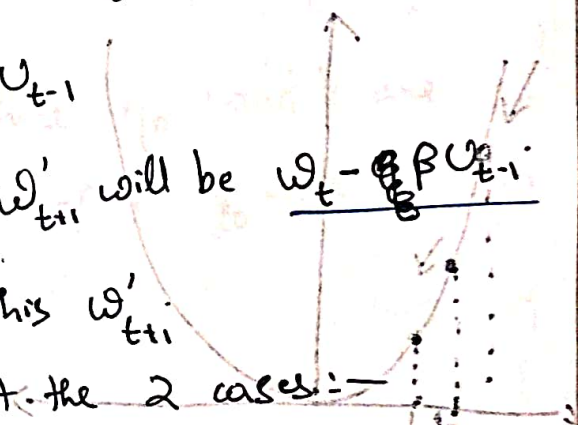
$\nabla w'_{t+1}$, now let's look at the 2 cases:-

1) Not overshooting:- if we don't overshoot, then $\nabla w'_{t+1}$ will be in line with the previous directions, so if v_t is calculated like this:-

$v_t = \nabla(w'_{t+1}) + \beta v_{t-1}$ (it would add on the momentum since both $\nabla w'_{t+1}$ and βv_{t-1} are similar directions).

actual $\Rightarrow w_{t+1} = w_t - \eta v_t$

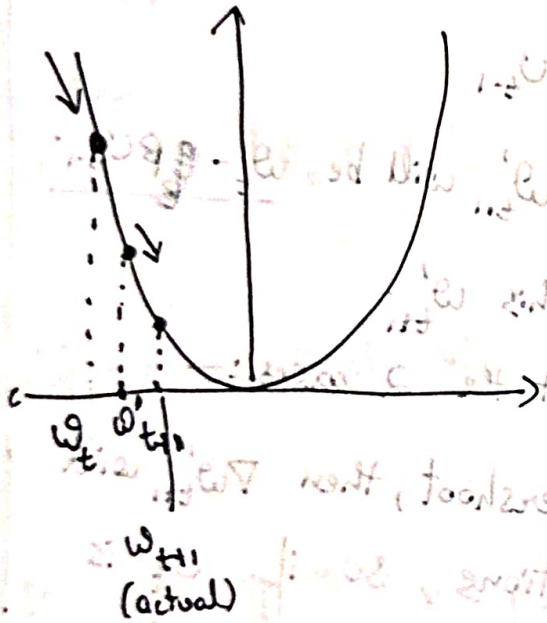
2) Overshoot: if we overshoot, then $\nabla w'_{t+1}$ will be in opposite like direction with the previous directions, so if v_t is calculated like this:-



$v_t = \nabla w'_{t+1} + \beta v_{t-1}$; it would reduce the momentum affect as $\nabla w'_{t+1}$ and βv_{t-1} are opposite like directions. Hence this will prevent/reduce overshooting as it would reduce the momentum affect.

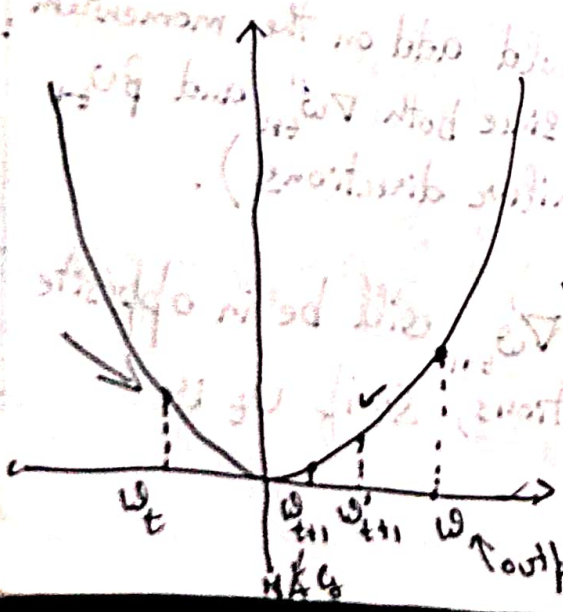
In both cases:-

$$v_t = \nabla(w_t - \beta v_{t-1}) + \beta v_{t-1}$$
$$w_{t+1} = w_t - \eta v_t$$



$w'_{t+1} = w_t - \beta v_{t-1}$ - looking ahead to w'_{t+1}

In this case, suggests that we should go along with the momentum and leads us to w_{t+1}



In this case looking ahead to w'_{t+1} suggests that we will overshoot and it cancels some momentum to reduce overshoot and leads us to w_{t+1}

w would have been the output for normal Momentum Based GD. (Overshooting is high)

↑ output for Momentum Based GD.

so NAG oscillates lesser than MGD.

One more way to speed up convergence is to set the correct learning rate η. Tips for Annealing the learning rate.

1) Halve the learning rate after every 5 epochs.

2) Halve the learning rate after an epoch if the validation error is more than what it was at the end of the previous epoch. So if the validation error is more at the end of current epoch, then reinitialize the weights to what was learned after previous epoch and run the current epoch again with $1/2$ the learning rate.

3) Exponential Decay :- $\eta = \eta_0 e^{-kt}$ where η_0 and k are hyperparameters and t is the step number. η_0 or $k \uparrow$ then the decay is more rapid.

4) $1/t$ decay, $\eta = \frac{\eta_0}{1+kt}$

5) For Momentum algorithms

$$\beta_t = \min\left(1 - 2^{-1-\log_2\left(\left\lfloor \frac{t}{250} \right\rfloor + 1\right)}, \beta_{max}\right), \text{ where}$$

β_{max} is chosen from $\{0.999, 0.995, 0.99, 0.9, 0\}$

6) Line Search try many values of η for each step, choose the one which causes the least loss.

All the methods discussed above change the learning rate for the whole gradient. This means irrespective of the features, the learning rate remains the same for all. But suppose we take the gradient

for the whole dataset and some of the features were important but very sparse. In this case the gradient update will be $w_{t+1} = w_t - \eta \nabla w_t$.

in the gradient

Naturally for sparse features, the corresponding element will be very small. Hence the updates made for these sparse features will be small as compared to other dense features. At the end of the algorithm (exhausting the iteration limit) the w^* for the sparse features may not be the optimal value as the updates were small in the iteration. Can we adapt the learning rate for each feature based on their frequency? (sparse features having a relatively higher learning rate).

AdaGrad :- Decay the learning rate for each parameter in proportion to their update history (more updates means more decay).

(larger magnitude)

Update Rule for AdaGrad :- $V_t = V_{t-1} + (\nabla w_t)^2$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{V_t + \epsilon}} * \nabla w_t$$

↑ scalar, as it refers to gradient w.r.t each parameter

$V_t = \sum_{k=0}^{t} (\nabla w_k)^2$

1) For dense features, V_t grows much faster as ∇w is large, so the decay in the learning rate is much faster.

2) For sparse features, V_t grows slowly as ∇w is small, so the decay in the learning rate is slow.

⇒ Adagrad slows down near the minimum due to decaying learning rate. This is because after some iterations, v_t is large and decay is too much, especially for dense features. Moreover, near the minima, the gradients are also small, which when combined with large accumulated v_t slows down the algorithm. v_t keeps on increasing.

⇒ To avoid this problem, we can use an algorithm that decays the denominator / v_t in such a way that in the steep areas, v_t is large but in gentle areas v_t decreases.

(effective learning is small) (effective learning is large)

This is Done in RMS Prop algorithm

$$v_t = \beta v_{t-1} + (1-\beta)(\nabla w_t)^2 \qquad 0 \le \beta \le 1, \text{ typically } 0.9, 0.99.$$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{v_t + \epsilon}} \nabla w_t$$

Let us expand v_t ($\beta = 0.9$)

$$v_3 = 0.9 v_2 + 0.1 \nabla w_3^2$$

$$v_3 = 0.9(0.9 v_1 + 0.1 \nabla w_2^2) + 0.1 \nabla w_3^2$$

$$v_1 = 0.9 v_0 + 0.1 \nabla w_1^2$$

$$v_0 = 0.1 \nabla w_0^2$$

$$v_1 = 0.1 \times 0.9 w_0^2 + 0.1 \nabla w_1^2$$

$$\Rightarrow v_3 = 0.9\left(0.9\left(0.1 \times 0.9 \nabla w_0^2 + 0.1 \nabla w_1^2\right) + 0.1 \nabla w_2^2\right) + 0.1 \nabla w_3^2$$

$$v_3 = 0.9\left(0.1 \times 0.9^2 \nabla w_0^2 + 0.1 \times 0.9 \nabla w_1^2 + 0.1 \nabla w_2^2\right) + 0.1 \nabla w_3^2$$

$$= 0.1 \times 0.9^3 \nabla w_0^2 + 0.1 \times 0.9^2 \nabla w_1^2 + 0.1 \times 0.9 \nabla w_2^2 + 0.1 \nabla w_3^2$$

$$= 0.1 \sum_{i=0}^{3} \beta^{3-i} (0.9) \nabla w_i$$

$\therefore v_t = (1-\beta)\sum_{i=0}^{t} \beta^{(t-i)} (\nabla w_i)^2$, we see that as t increases, the contribution from the early gradients decays to 0.

In steep areas ∇w's are large hence v_t also becomes large and the learning rate reduces.

In gentle areas, ∇w's are small and v_t decays hence the learning rate increases.

So RMSProp retains the advantages of AdaGrad for sparse and dense features but also has better convergence behaviour because the learning rate can increase or decrease.

One problem RMSProp faces is that it can oscillate around the solution. Near the minima ∇w_i are small hence making v_t decay to 0. (0)

This makes the effective learning rate constant $\frac{\eta}{\sqrt{\epsilon}}$, which is susceptible to oscillation. So RMSProp depends on the initialization of these parameters.

⇒ To reduce this effect we can adapt the numerator as well this allows us to avoid the dependency on the initializations. This algorithm is called AdaDelta. The algorithm is as follows:-

$$v_t = \beta v_{t-1} + (1-\beta)(\nabla w_t)^2$$

$$\Delta w_t = -\sqrt{\frac{u_{t-1}+\epsilon}{v_t+\epsilon}}\,\nabla w_t$$

$$w_{t+1} = w_t + \Delta w_t$$

$$u_t = \beta u_{t-1} + (1-\beta)(\Delta w_t)^2$$

Notice that Δw_t depends on current v_t and previous u_{t-1}.

1) In steep areas v_t is more updated and increases. but u_{t-1} is not yet updated, so in Δw_t numerator is smaller than the denominator, hence the learning rate is smaller.

$u_t < v_t$

2) In flat regions, v_t is more update and reduces, but v_{t-1} is not yet updated, so in Δw_t, the denominator is small, the learning rate becomes larger. ($v_t > v_t$)

⇒ AdaDelta adapts the learning rate more smoothly than RMSProp algorithm.

Adam (Adaptive Moments): Very similar to the RMSProp algorithm, but introduces momentum in the calculation. So instead of directly using ∇w_t in the gradient update, we update it using the momentum term

$$m_t = \beta_1 m_{t-1} + (1-\beta_2)\nabla w_t$$

Scaling/removing bias from m_t ⇒ $\hat{m}_t = \frac{m_t}{(1-\beta_1^t)}$ (β_1^t to the power t)
(explained later)

$$v_t = \beta_2 v_{t-1} + (1-\beta_1)(\nabla w_t)^2, \text{ then correcting the bias: } \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

update rule: $w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}}\hat{m}_t$

In this momentum, we give more weightage to overall previous gradients in the form of $\beta_1 m_{t-1}$, the current gradient is given slightly less weightage since it gets multiplied by $(1-\beta_1)$. It is like a moving exponential average of ∇w_t's. We can't calculate $E[\nabla w]$, as it will be expensive, but what about

$$E[m_t] = E\left[(1-\beta_1)\sum_{t=0}^{t}\beta_1^{(t-i)}\nabla w_i\right]$$

$$= (1-\beta_1)\sum_{i=0}^{t}\beta_1^{(t-i)}E[\nabla w_i]$$

Since $E[\nabla w_i] = E[\nabla w]$

$$\Rightarrow E[m_t] = (1-\beta_1) \sum_{i=0}^{t} \beta^{t-i} \times E[\nabla w]$$

$$= (1-\beta_1) E[\nabla w] \sum_{i=0}^{t} \beta^{(t-1)}$$

$$= (1-\beta_1) E[\nabla w] \times \frac{(1-\beta_1^t)}{(1-\beta_1)}$$

$E\left[\frac{m_t}{1-\beta_1^t}\right] = E[\nabla w]$. therefore $\frac{m_t}{1-\beta_1^t}$ gives an unbiased estimate of ∇w.

⇒ A lack of initialization bias correction would lead to initial steps that are much larger. Bias correction also ensures that initial learning rate is not very huge.

⇒ One more advantage of bias correction is that it reduces the affect of noisy points taken in Mini batch GD for gradient computation.

Max Prop :- If we closely revist the formula for RMS Prop:-

$$V_t = \beta V_{t-1} + (1-\beta)(\nabla w_t)^2 \quad ; \quad w_{t+1} = w_t - \frac{\eta}{\sqrt{V_t + \epsilon}} \nabla w_t$$

Although it achieves the requirement of adjusting the learning rate based on the gradient. But is still aggressive and that makes it susceptible to noise. Take for example a sparse feature in a Mini Batch G.D and the gradient is considerable. Now a batch comes which was noisy and no data point had this features, so gradient will be zero, in the current rule, the zero gradient which is noisy will affect V_t and reduce it and hence increase the learning rate. This is undesirable.

⇒ We change the way V_t is calculated to make it less susceptible to noise. That is done in the max prop, instead of using the

L^2 norm (squares) we can use the max norm. So the update formula becomes:-

$$v_t = \max\{\beta v_{t-1}, |\nabla w_t|\}$$

$$w_{t+1} = w_t - \frac{\eta}{v_t + \epsilon} \nabla w_t$$

↳ square root is removed as max norm is used now.

Thinking about the same situation, zero gradient will never be picked. But still the v_t decays as β is multiplied, so if the behaviour is consistent, v_t will decay and the original behaviour is still retained.

~~But in the case of~~

Introducing Max norm in Adam :- AdaMax

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)(\nabla w_t), \quad \hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

$$v_t = \max\{\beta_2 v_{t-1}, |\nabla w_t|\}, \quad w_{t+1} = w_t - \frac{\eta}{v_t + \epsilon} \hat{m}_t$$

notice that $\hat{v}_t$ is not needed as max norm does not drastically increase the learning rate.

⇒ Introducing Nesterov Accerelation in Adam :- NADAM

Slightly different notation (did not understand)

$$m_{t+1} = \beta_1 m_t + (1-\beta_1)\nabla w_t$$

$$\hat{m}_{t+1} = \frac{m_{t+1}}{1-\beta_1^{t+1}}$$

$$v_{t+1} = \beta_2 v_t + (1-\beta_2)(\nabla w_t)^2 \qquad \hat{v}_{t+1} = \frac{v_{t+1}}{1-\beta_2^{t+1}}$$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_{t+1}} + \epsilon}\left(\beta_1 \hat{m}_{t+1} + \frac{(1-\beta_1)\nabla w_t}{1-\beta_1^{t+1}}\right)$$

Learning rate Schemes

Based on epochs
1) Step Decay
2) Exponential Decay
3) Cyclical
4) Cosine Annealing

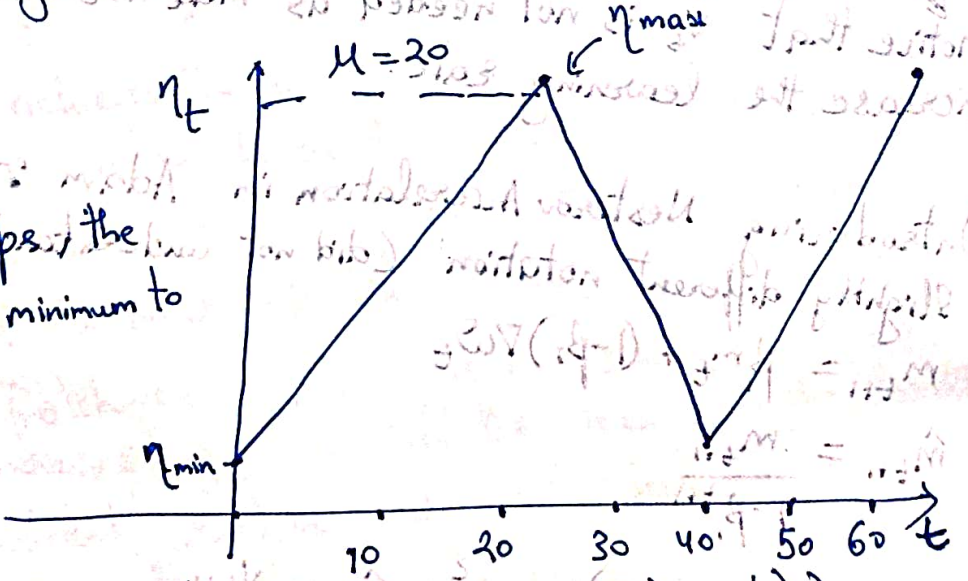
Base Validation
1) Line Search
2) Log Search

Based on Gradients
1) AdaGrad
2) RMSProp
3) AdaDelta
4) Adam
5) AdaMax
6) Nadam
7) AMSGrad
8) AdamW.

Cyclical Learning Rate :- At saddle points, the gradients become small, and it becomes difficult to move out from those areas. If we allow the learning rate to increase after some iterations irrespective of the gradient, then there is a chance to escape the saddle points. Adaptive learning rate helps in escaping saddle points but they are computationally expensive. We can vary learning rate cyclically as a simple strategy.

i) Triangular :-

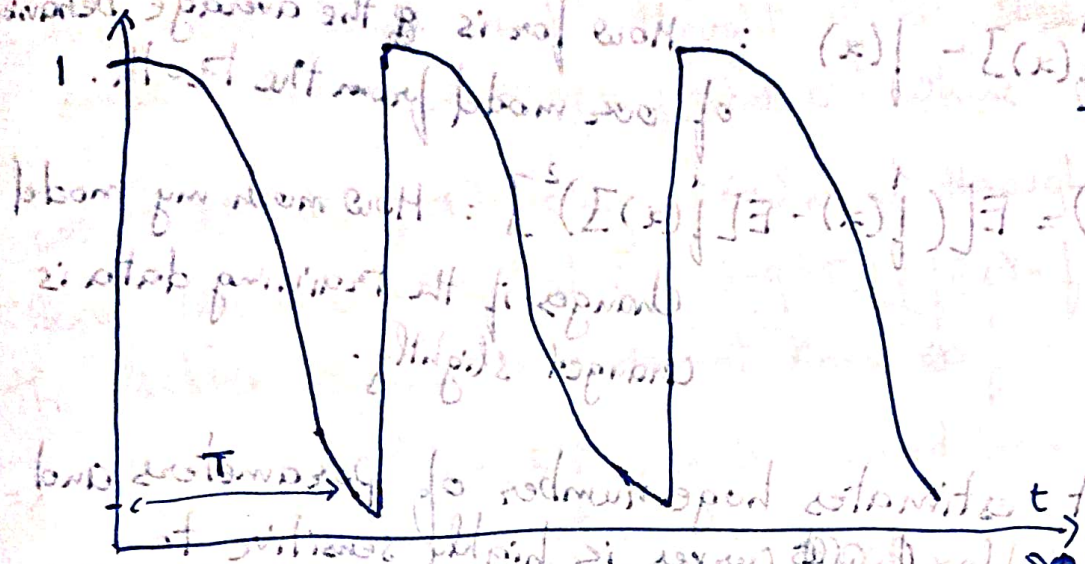
After every 20 steps, the learning rate reaches from minimum to maximum.



Formula is given by:-

$$\eta_t = \eta_{min} + (\eta_{max} - \eta_{min}) \cdot \max\left(0, \left(1 - \left|\frac{t}{U} - \left(2\left\lfloor 1 + \frac{t}{2U} \right\rfloor\right) + 1\right|\right)\right)$$

ii) Cosine Annealing (Warm-restart)



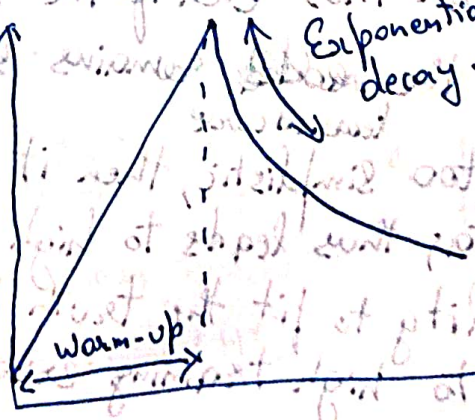
T is the restart interval.

$$\eta_t = \eta_{min} + \frac{\eta_{max} - \eta_{min}}{2}\left(1 + \cos\left(\pi \frac{t\%(T+1)}{T}\right)\right)$$

But due to abrupt increases in the learning rate, it may escape (overshoot) near the minima. In that case we could use techniques such as early stopping to roll back to the minimum.

Warm Start (found to be helpful in transformer architectures)

Typically, we set the initial learning rate to a high value and then decay it. On the contrary, using a low initial learning rate helps the model to warm and converge better. This is called warm start. This is because, initially the weights are random and we allow the model to make small steps in early gradient directions. Then after certain Number of iterations, the usual exponential decay strategy is applied. Number of warm-up rounds can be a choice.



Bias and Variance of a Model :-

Bias :- $E[\hat{f}(x)] - f(x)$: How far is the average behaviour of our model from the truth.

Variance $(\hat{f}(x)) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$: How much my model changes if the training data is changed slightly.

A model that estimates huge number of parameters and capable to fit complex curves, is highly sensitive to training data. If this model is trained on a slighly different sample of training data, it will significantly differ in its learned parameters. Thus it has high variance. Since this model can fit the training data, it leads to low errors and hence low bias as it has the capability to fit the truth closely. But sometimes, if the data has noise, then this model gets severely affected because even though the training error will be small as always, but, on unseen data, it performs poorly. This is called overfitting. It still has low bias (usually) as the model is capable to fit the truth if we ignore the noise. So overfitting is a consequence of high variance and noise.

⇒ Similarly a model that estimates very few number of parameters. Thus even if the training data sample is slighly changed, the learned curve remains similar. Thus low variance. If the model is too simplistic, then it may fail to capture the structure in the data, thus leads to high bias as the model does not have the capability to fit the truth irrespective of training data. This leads to high training error and is called underfitting.

Estimating error from test data

the true labels can be thought of as a combination of a function and some noise. So: $y = f(x) + \epsilon$ where $\epsilon \in N(0, \sigma^2)$

we estimate $f(x)$ using $\hat{f}(x)$ or $\hat{y}$ through our model. Hence we are interested in using $E[(\hat{f}(x) - f(x))^2]$. But the problem is that we do not know f.

$\Rightarrow E[(\hat{f}(x) - f(x))^2]$ can be derived from the following formula:

$$E[(\hat{y} - y)^2] = E[((\hat{f}(x) - f(x)) - \epsilon)^2]$$

$$= E[(\hat{f}(x) - f(x))^2] + E[\epsilon^2] - 2E[\epsilon(\hat{f}(x) - f(x))]$$

$$\Rightarrow E[(\hat{f}(x) - f(x))^2] = E[(\hat{y} - y)^2] + 2E[\epsilon(\hat{f}(x) - f(x))] - E[\epsilon^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 - \sigma^2 + 2E[\epsilon(\hat{f}(x) - f(x))]$$

For 2 independent variables X and Y, and one having 0 expectation, $E[XY] = E[X] \times E[Y] = 0$ (i.e. either $E[X]$ or $E[Y]$ is 0)

Since in test data, none of the points participated in estimating $\hat{f}(x)$, $(y - f(x))$ is independent from $(\hat{f}(x) - f(x))$. Hence ϵ is independent from $(\hat{f}(x) - f(x))$.

So $E[(\hat{f}(x) - f(x))^2] = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 - \sigma^2$ ← depends on the noise.

If the data is sufficiently clean, then this is a good estimate of the true error (LHS).

If instead we used training data for estimation then ϵ is not independent from $\hat{f}(x) - f(x)$. Hence $E[\epsilon(\hat{f}(x) - f(x))] \neq 0$

the true error LHS is larger than the empirical error.

$$\frac{1}{n}\sum(\hat{y}_i - y_i)^2.$$

So using the empirical training error does not give a true representation of the actual error. But training is based on minimizing this error.

Stein's Lemma:- $\frac{1}{n}\sum \epsilon_i(\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n} \frac{\partial \hat{f}(x_i)}{\partial y_i}$

If $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ is large when there is high variance in the model, changing y_i brings large change in the learned model.

LHS of Steins Lemma is similar to $E[\epsilon(\hat{f}(x_0) - f(x))]$

So we can say that true error is a function of model complexity.

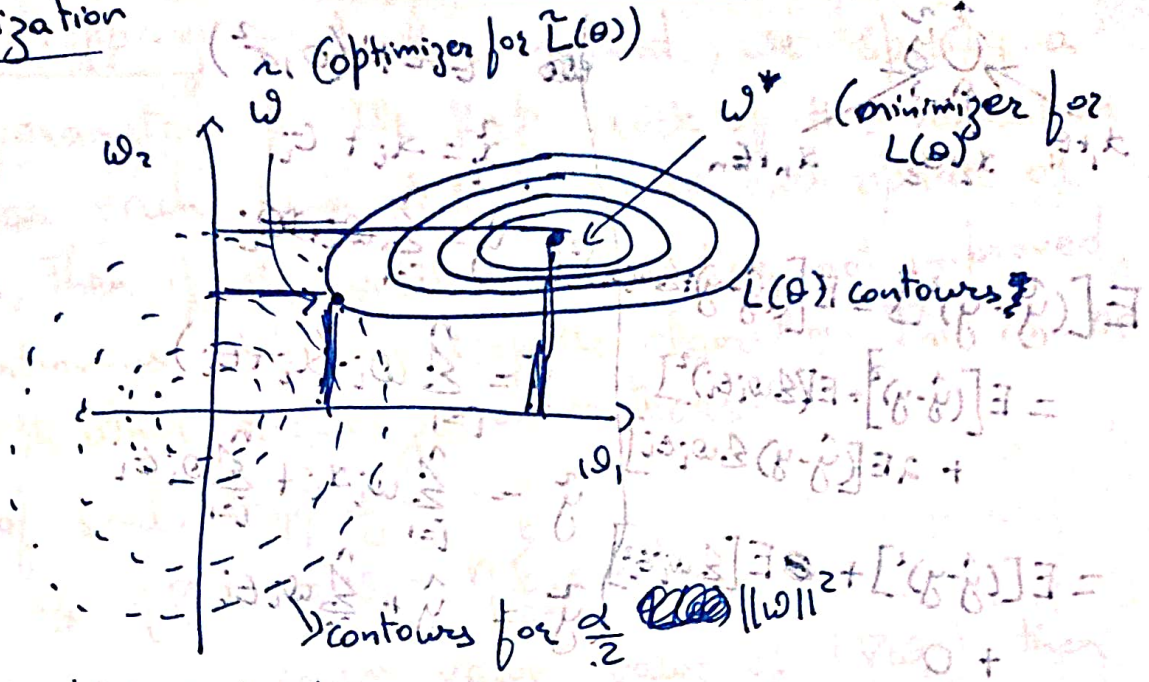
$$E[(\hat{f}(x) - f(x))^2] = \text{empirical train error} - \text{small constant} + \Omega(\text{model complexity})$$

The more complex the model, the true error may increase.

So it is better to minimize a modified objective function that also incorporates model complexity into it.

$$\min_{\theta} \hat{L}(\theta) \rightarrow \left(L_{train}(\theta) + \Omega(\theta)\right).$$ But instead of calculating $\Omega(\theta)$ as given by steins lemma we can calculate it using an approximation that captures model complexity. This strategy is called regularization.

i) L2 Regularization



of Hessian of $L(\theta)$

the eigen value, corresponding to w_1 will be smaller. Hence the ellipse is elongated in w_1 axis. Moreover, there is more shrinkage in w_1 as compared to w_2. This can also be inferred through the contours of $L(\theta)$, because in the axis of w_1, the contours are further apart, so derivative is small. So Loss is less sensitive to w_1 than w_2. So w_1 is shriked more.

ii) Data Augmentation :- we can introduce new training data points by slightly perturbing the already present data points and then keeping their labels same. Example, shifting or rotating an image to create a new image.

iii) Injecting noise at inputs :- at each epoch, corrupt some part of the input, thus it becomes difficult for the model to overfit on the training data as at each epoch the training data slightly changes. Can be thought of data augmentation at each epoch. For an input output network, without any hidden layers, the noise injection at input is equivalent to L2 regularization

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

$\hat{x}_i = x_i + \epsilon_i$

$\hat{y} = \sum_{i=1}^{n} \omega_i \hat{x}_i$

$\hat{y} = \sum_{i=1}^{n} \omega_i (x_i + \epsilon_i)$

$\hat{y} = \sum_{i=1}^{n} \omega_i x_i + \sum_{i=1}^{n} \omega_i \epsilon_i$

$\hat{y} = \hat{y} + \sum_{i=1}^{n} \omega_i \epsilon_i$

$E[(\hat{y}-y)^2] = E[((\hat{y}-y)+\sum \omega_i \epsilon_i)^2]$

$= E[(\hat{y}-y)^2] + E[(\sum \omega_i \epsilon_i)^2]$

$+ 2E[(\hat{y}-y)\sum \omega_i \epsilon_i]$

$= E[(\hat{y}-y)^2] + E[\sum \omega_i^2 \epsilon_i^2]$

$+ 0$

This is because $E[(\sum \omega_i \epsilon_i)^2] = E[\sum \omega_i^2 \epsilon_i^2]$

Since ϵ_i is independent from ϵ_j and $E[\epsilon_i \epsilon_j] = E[\epsilon_i]E[\epsilon_j]$

$= 0$

So, all interacting terms in LHS will be 0.

Hence only square terms will survive.

Also ϵ_i is independent from $(\hat{y}-y)$ so $E[(\hat{y}-y)\epsilon_i \omega_i]$

$= E[\hat{y}-y] \cdot \omega_i E[\epsilon_i]$

$\hookrightarrow 0$

$= 0$

so $E[(\hat{y}-y)^2] = L(\theta) + E[\sum \omega_i^2 \epsilon_i^2]$

$$\boxed{E[(\hat{y}-y)^2] = L(\theta) + \sigma^2 \sum \omega_i^2}$$ as $E[\epsilon_i^2] = \sigma^2$

iv) Injecting noise at the outputs :- Instead of fitting on the true labels, we perturb the true labels by adding some noise. This has a similar effect to L2 regularization.

v) Early Stopping :- In this method, we specify a patience parameter p. We keep track of the validation error after each step. A step can be each update or each epoch. Then if the validation error has not improved after p continuous steps we stop the algorithm. This can be combined with other regularization methods.

Analysis of Early Stopping :-

$$w_{t+1} = w_t - \eta \nabla w_t = w_0 - \eta \sum_{t=0}^{t} \nabla w_i$$

Let T be the maximum (absolute) value of $|\nabla w_i|$, then

$$|w_{t+1} - w_0| \leq \eta T (t+1)$$

, thus t controls the upper bound of $|w_{t+1} - w_0|$. If we limit t, in some sense we are limiting $|w_{t+1} - w_0|$ and if $w_0 = 0$, then this acts as a regularizer on the weights

vi) Ensemble Methods :- bagging does not help if the ensembles are correlated as they will make similar mistakes. Bagging is most effective when ensemble models are decorrelated as mistake of one rectified by the other model

2 options available in bagging are :-

i) Train different neural networks having different architecture

ii) Train multiple instances of the same network using different training samples.

Both these options are computationally expensive.

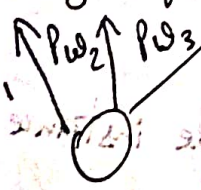
vii) Dropout :- In this method we train a single instance of neural network, but at each step (update or epoch) we drop some neurons by temporarily removing all their incoming and outgoing edges. Neurons to drop are decided randomly

sampled, where some probability is associated with each neuron. For example each neuron has $p = 0.8$ probability of being retained. The total number of possible neural networks that can be made are exponential in the number of total nodes. The full algorithm is given below:-

1) Initialize the network by choosing which neurons are active and participate in the computation.
2) Do forward propogation using the weights from previous iterations : updated
3) Do backward propogation and update the weights that are active
4) Repeat the procedure. We can decide if we want to keep one network for 1 update or 1 epoch. After each step, a new or in rare cases a previous configuration is made active and weights are updated.

Each network is trained twice rarely, but each weight may be updated many times in different configurations.
subset

At test time, the output is calculated from the whole network but the output of each neuron will be multiplied by the probability it was retained with in each sample which is p. or by the fraction of times it was active during training

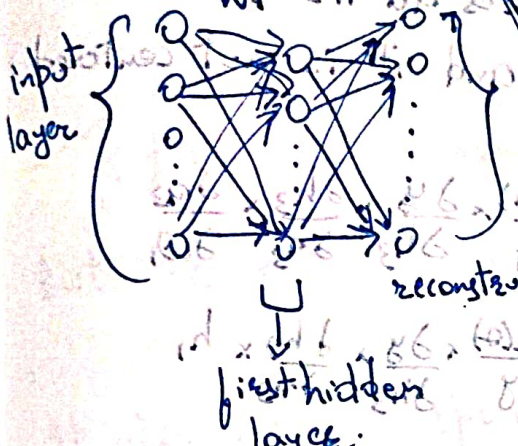


Dropout essentially applies a making noise to the hidden units. It prevents units from Co adapting. A hidden unit cannot rely too much on other units as they may get dropped out any time. It allows the model to discover new features and makes it more robust.

Unsupervised Pre-Training :-

Instead of initializing the parameters randomly, is there a way to initialize them?

Consider the first hidden layer, having some n_1 neurons.



If we ignore the rest of the network and set up a new network such that task of this network is to reconstruct the input layer with the least error possible, with only the 1 hidden layer consisting of the neurons from 1st hidden layer of the original network:

So the objective is such that:- $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_0}(\hat{x}_{ij}-x_{ij})^2$

each input is $\mathbb{R}^{n_0}$, and there are n such inputs. The reconstructed output is also $\mathbb{R}^{n_0}$. This is like PCA, the objective is same.

Whatever weights are learned as W_1 (between input to 1st hidden layer) can be used as initialization in the original network.

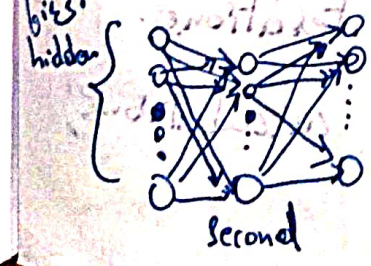
For the weights of the second layer (W_2) we calculate the output of each neuron in the first layer for each data point using the weights in the first step. So output of the 1st hidden layer will be $\mathbb{R}^{n_1}$, and there are n such inputs.

The objective will now be:- $\min \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_1}(\hat{h}_{ij}^{(1)}-h_{ij}^{(1)})^2$

$h_{ij}^{(1)}$ is the output from the j^{th} neuron of the 1st layer for the i^{th} point.

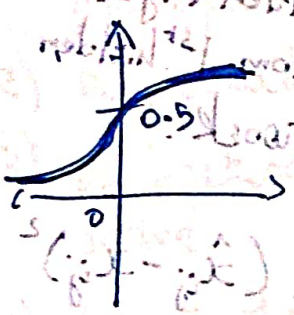
From this we obtain W_2.

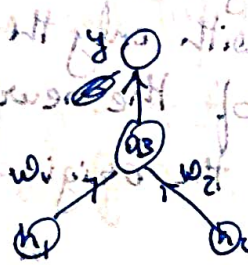
We do this to obtain W_1 and then initialize these as the weights for the original network.



Activation Functions :- 1) Linear :- If we have a deep linear neural network, then it faces difficulty in approximate all the functions. Therefore some non-linearity is needed.

2) Sigmoid :- Sigmoid function compresses all its domain into $[0,1]$. The output of sigmoid is positive and it is not centered around 0.



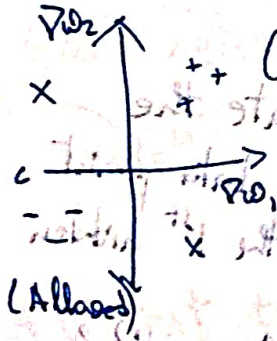


$$\nabla w_1 = \frac{\partial L(\theta)}{\partial y} \times \frac{\partial y}{\partial h_3} \times \frac{\partial h_3}{\partial a_3} \times \frac{\partial a_3}{\partial w_1}$$

$$= \frac{\partial L(\theta)}{\partial y} \times \frac{dy}{dh_3} \times \frac{dh_3}{da_3} \times h_1$$

$$\nabla w_2 = \frac{\partial L(\theta)}{\partial y} \times \frac{dy}{dh_3} \times \frac{dh_3}{da_3} \times h_2$$

Since neurons are sigmoid, h_1 and h_2 are positive. Besides h_1 and h_2, ∇w_1 and ∇w_2 are same. Hence both ∇w_1 and ∇w_2 will have the same sign. Second and fourth quadrant gradients cannot be easily calculated as sign of the components are opposite.



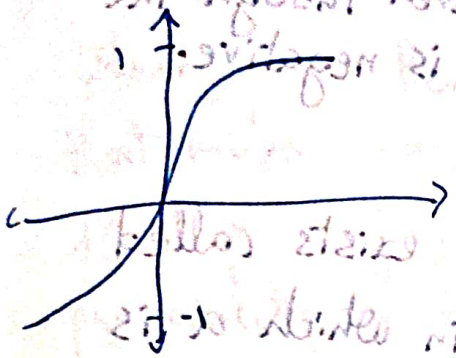
0 centering removes this problem.

One more problem with sigmoid neuron is that near large absolute values of input, the neuron outputs 0 or 1, but the gradient vanishes as the rate of change is very small. So incorrect initializations of weights makes the neuron saturate (0, 1) and after saturation, since the gradient is small, the updates happen very small which keeps it saturated in the next iterations.

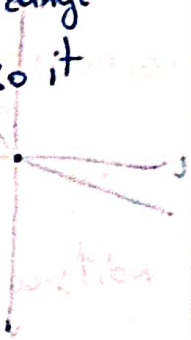
Third problem is that sigmoid involves sigmoid calculations, which is also expensive.

3) Tanh function : $\tanh(x)$

$$\frac{d}{dx}\tanh(x) = 1 - \tanh^2(x)$$



Compresses all its input to the range $[-1, 1]$. It is zero centered so it removes the opposite sign gradient.



But the problem of vanishing gradient still exists. and saturation.

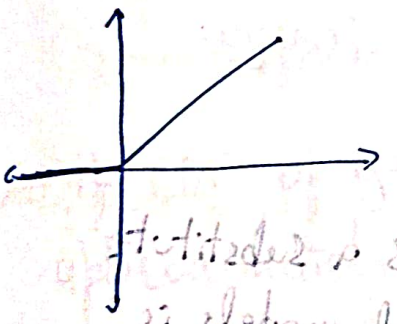
It also involves exponential computations, so it is also relatively expensive.

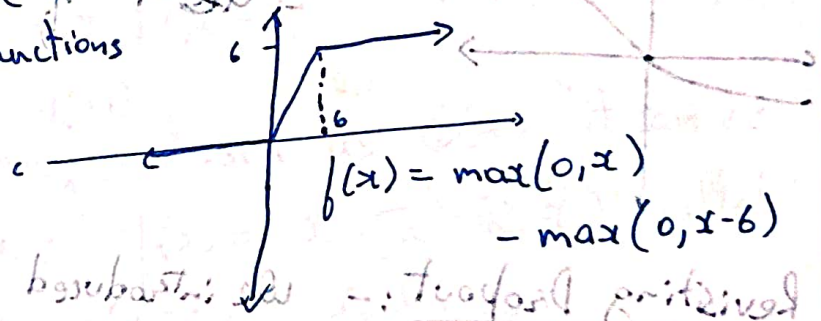
4) Relu (Rectified Linear Unit)

$$f(x) = \max(0, x)$$

Linear functions

We can recover piecewise





Advantages of ReLU:

Does not saturate in the positive region, computationally efficient, in practice it converges much faster than sigmoid / tanh

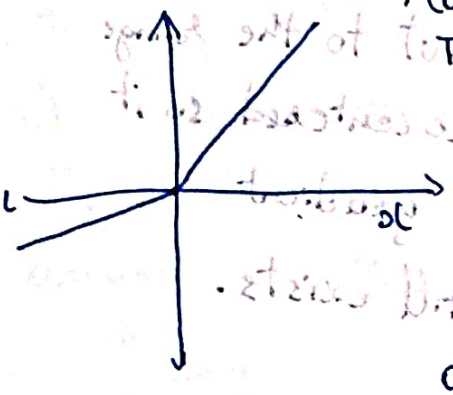
$$\frac{d}{dx}\text{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

=) If bias of some ReLU is set to a large negative value, $b << 0$, then the output as well as gradient for this neuron will be 0. Hence the vanishing gradient problem still exists for one half of input.

Leaky ReLu : No Saturation.

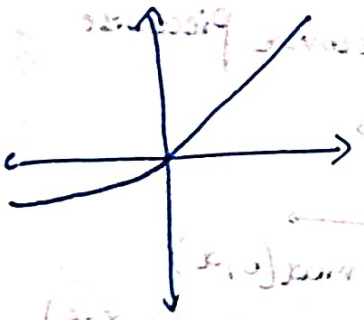
$f(x) = \max(0.1x, x)$ → can be α, any $\alpha > 0$. Typically small.

Some gradient value still exists even though the input is negative.



Another version exists called Parametric ReLu in which α is also treated as a parameter, which gets updated during backpropagation.

Exponential Linear Unit



$f(x) = x$ if $x > 0$

$= \alpha e^x - 1$ if $x \leq 0$

Revisiting Dropout :- We introduced Dropout as a substitute to bagging. But in bagging, each individual models is overfit on the its corresponding bagged training data. But in dropout, each subnetwork configuration hardly fits on the data. To achieve the bagging effect with dropout, we introduce MaxOut neuron. $\Rightarrow \max\{w_1 x + b_1, \cdots, w_n x + b_n\}$. Only the neuron that contributes the most gets the update, as for others, gradient is 0.

Dropouts can also be written as $U(x) = \begin{cases} 1.x & , p \\ 0.x & , (1-p) \end{cases}$

output of a neuron is as it is (x) if it got selected, the probability of that is p. And output is 0 otherwise.

It makes sense to combine and represent more compactly:- $f(x) = m \cdot x$

where $m \sim \text{Bernoulli}(\phi(x))$ where $\phi(x)$ is a function that maps $-\infty, -\infty$ between $[0, 1]$. For dropout $\phi(x) = p$ works, as each neuron will be selected with probability of p.

Alternatively it can be a cummulative distribution function of a Gaussian variable or even logistic function.

$$E[f(x)] = E[m \cdot x]$$
$$= x\, E[m]$$
$$= x \left[1 \times \phi(x) + 0 \times (1 - \phi(x))\right]$$
$$E[f(x)] = x\,\phi(x) \rightsquigarrow P(X \le x)$$

if CDF of gaussian function, $N(0,1)$ is chosen as $\phi(x)$, it can be approximately written as $f(x) = x\,\sigma(1.72x)$ called GeLU.

GeLU. $\text{GeLU}(x) = x \cdot \sigma(1.72x)$

SELU (Scaled Exponential Linear Unit): Linear unit with output scaled such that mean is 0 and unit variance.

$$f(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha e^x - \alpha & \text{if } x \le 0 \end{cases}$$

$\lambda \approx 1.0507009$
$\alpha \approx 1.6773263$

Studies found that functions of the form $f(x) = x\,\sigma(\beta x)$ are good. Function of such forms where β is a learnable parameter is called SWISH. if $\beta = 1.702$, it is GeLU.

when $f(x) = x\sigma(x)$. it is called SILU, Sigmoid weighted linear unit.

Convolution operation:-

$$S_{ij} = (I * K)_{ij} = \sum_{a=\lfloor -m/2 \rfloor}^{\lfloor m/2 \rfloor} \sum_{b=\lfloor -n/2 \rfloor}^{\lfloor n/2 \rfloor} I_{i-a, j-b} \cdot K_{\frac{m}{2}+a, \frac{n}{2}+b}$$

3D-filter is referred to as a volume. In case of 3-D filters, the volume is slid over the 3D image same as 2D case, and the convolution result is 2D.

Defining Quantities :-

⇒ Width (W_1), Height (H_1) and Depth (D_1) of the original input

⇒ The Stride S (increment in the slide of the kernel)

⇒ The number of filters is K.

⇒ The spatial extent (F) of each filter (the depth of each filter is same as the depth of each input)

⇒ The output is $W_2 \times D_2 \times D_2$.

Usually the extent of the filter is an odd integer as center of the matrix is properly defined.

If no padding is applied, the convolution reduces the size of the image as the pixels in the border of the input, convolution is not defined for.

The formula for output dimensions is given as:-

$$W_2 = W_1 - (F_1 - 1) = W_2 - F_1 + 1$$

$$H_2 = H_1 - F_1 + 1$$

This is because if the $F = 2a+1$ for $a \geq 0$, a columns are lost in the left border and a colums are lost in the right border of the input.

So total colums lost $= 2a = F-1$

$\therefore$ Remaining $= W_1 - (F-1)$

Similarly for height, remaining height $= H_1 - (F-1)$

If we add padding of P pixels around the image, then the formula is upated as follows:-

$$W_2 = W_1 - F + 1 + 2P$$

$$H_2 = H_1 - F + 1 + 2P$$

Generally we do a stride $S = 1$ which is look at all pixels. $S = 2$ means that place kernel at every alternate pixel. (in both x and y dimension)

$$W_2 = \frac{W_1 - F + 2P}{S} + 1$$

$$H_2 = \frac{H_1 - F + 2P}{S} + 1$$

$\Rightarrow$ K filters will give us $K = 2D$ outputs. Hence $D_2 = K$

Convolutional Neural Networks :-

In traditional Machine learning, we would convert the image into a flat vector and the apply the classification algorithm. We can also apply edge detection algorithms and kernels before applying the classification algorithms. But these kernels are hard coded before the algorithm, and is treated completely as preprocessing.

We can also learn these kernels and treat them as params of the model. Not only this, multiple kernels can be learnt. Furthermore the neural network's architecture allows to learn multiple layers of kernels to produce the final output.

Difference between a CNN and a normal feedforward Network

In a feed-forward network, each neuron is connected to each neuron in the previous network. It is dense network and all weights are independently treated.

But in a CNN:-

Suppose there is an image of 4x4 given as input to the network, and a 2x2 kernel is trying to be learnt for the first hidden layer.

i) The input image is flattened :-

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

$\Rightarrow$ | 1 | 2 | . . . . | 15 | 16 |

Hence input layer has 16 neurons.

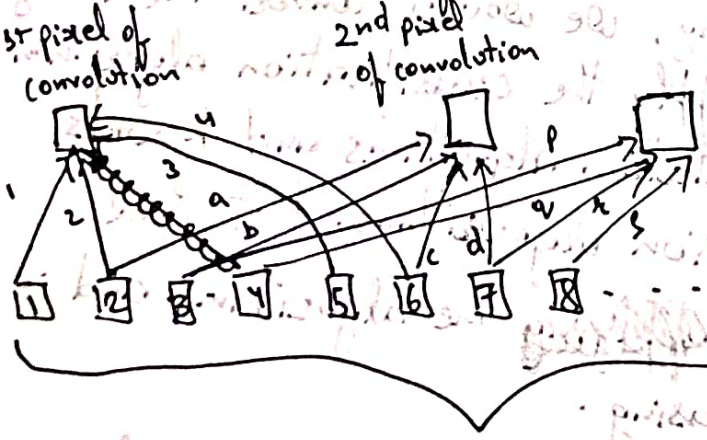
ii) The kernel is flattened :-

1	2
3	4

$\Rightarrow$ | 1 | 2 | 3 | 4 |

iii) The output image is calculated and flattened as well. So each pixel in the output of the convolution is a neuron in the next layer.

We can make the following observations :-



i) The output layer is sparsely connected with the previous layer. This is because the first pixel of convolution output is calculated using only 1, 2, 5 and 6 pixel of input. Similarly for all pixels. So the connectivity is sparse

ii) Since a single kernel is applied here, $1 = a = p$, $2 = b = q$, $3 = c = r$, $4 = d = s$

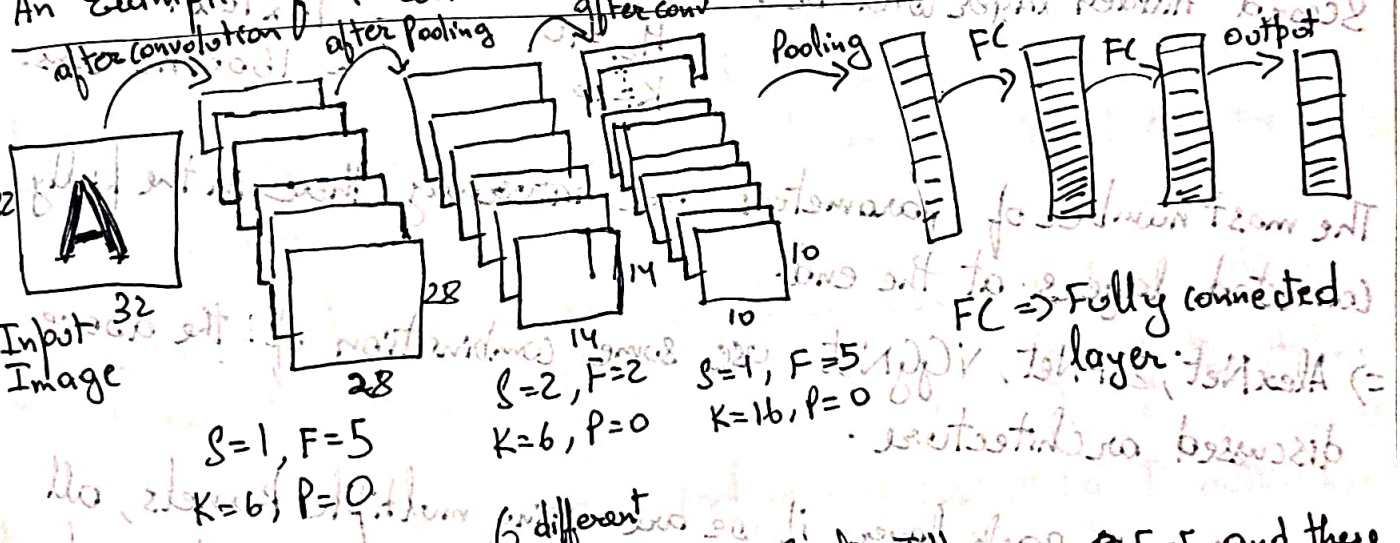
There is weight sharing in this neural network. More kernels can be applied but that also increases the number of output neurons (next convolution output, addition in depth) so weight sharing is still there.

Due to this weight sharing :- i) we look at the neighborhood of each pixel.

ii) We don't really loose information as in the next layers, there will be interaction between pixels which dependent on the input pixels. So indirect iteraction is there.

=> After performing convolution the result can be passed through an activation function.

An Example of a convolutional Neural Network:-



The first layer consists of 6 different 5x5 filters. Each Filter is 5x5 and there are 6 such such filters, so there are 6x25 = 150 parameters. Note that this means that there will be 150 "unique"/independent connections from the input layer to the first hidden layer. Moreover there are 28x28x6 neurons in the first hidden layers. This is decided by the formula discussed before.

The Layer after first hidden layer is the max pooling layer. In this layer is passed through a max filter independently to obtain Each component matrix in the depth of the first hidden the 6 corresponding matrices. Notice that a stride of 2 is used in the example, dimension rules apply in the same manner.

In max filter, just the max is selected within the neighborhood defined by the kernel.

Max pooling is done to reduce the size of the layers and hence reduce the number of parameters, preventing overfitting.

Also note that the max pooling layer learns no parameters, it is a fixed process involving no learnt parameters.

=) The next hidden layer (second) is again a convolutional layer with filter size of $5 \times 5 \times 6$. There are 16 such filters. Hence the number of parameters are 2,400. The number of neurons in the second hidden layer will be :-

$W_2 = 14 - 5 + 1 = 10$

$H_2 = 10$

$K = 16$

$= 10 \times 10 \times 16$

$= 1600$ neurons.

The most number of parameters are naturally there in the fully connected layers at the end.

=) AlexNet, ZFNet, VGGNet use some combination of the above discussed architecture.

Notice that in each layer, if we are using multiple kernels, all those kernels are the same, both in regard of type of kernel and size of kernel. We can combine these options and the concatenate them to produce the output.

But this introduces a problem, if a filter is $F \times F \times D$, then each entry in the output requires $F \times F \times D$ computations. This makes it difficult to apply multiple kernels. We can use 1×1 convolutions to reduce the depth of the input.

=) The $1 \times 1 \times D$ convolution operation applies across the depth of the input.

=) We can apply D_1 such 1×1 kernels to the input

image to produce a $W \times H \times D_1$ output, which is a reduction in the depth. Based on this convolved output we can use different sized kernels or pooling. Moreover we can also calculate another $W \times H \times D_2$ by 1×1 kernels and apply another set of kernels on this output. The possibilities are huge. Yes the computations are increased but the number of parameters in the later layers are greatly reduced.

To make the output consistent in dimensions, appropriate padding is added.

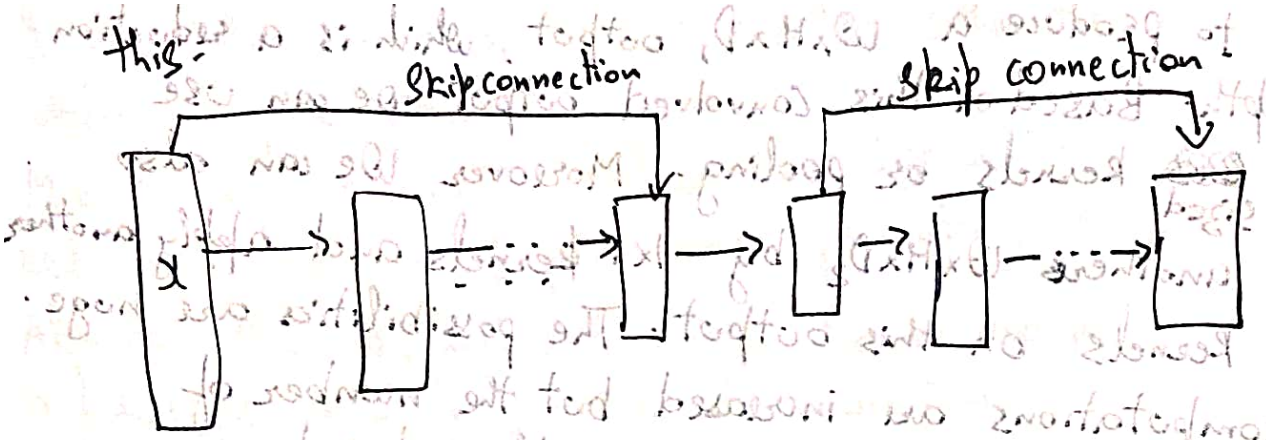
All these computations are done in parallel. This whole unit is called an inception module. The unit consists of 1×1, convolutions and pooling and other size convolutions. The output is concatenated and treated as the output of the inception module as a single layer.

=> One more change can be to do average pooling in the last layer before the fully connected layer. So if it has $E W \times F \times D$ output, then average pooling reduces it to D outputs, applied by taking average of each component matrix. This reduces the number of parameters.

All these improvements were used in the GoogleNet.

ResNet :- Suppose we have a shallow Network that performs well. But if we make it deeper, we expect it to perform at least as better as the shallow network. But it does not happen in practice. To improve the performance of the deep network, we can use skip connections. Skip connections supply a copy of input to further layers. We can introduce skip connections after every few layers. Infact, ResNet did

this:



Ways to Visualize and Interpret a CNN

=> For each neuron or set of neurons, we can feed in images to this CNN and identify the images which cause these neurons to fire.

=> We can then trace back to the path in the image which causes these neurons to fire.

This will allow us to understand which neuron(s) are excited at what input.

=> One way to do this is to plot the kernels / filters as images and interpret them as patterns. The filters detect the patterns observed in their image plot. The problem with this approach is that the results are interpretable only for the first layer. Because after that the filters are applied on the input of previous layers.

=> second approach is through occlusion experiments. In this method, we occlude (gray out) different patches in the image and see the effect on the predicted probability of the correct class. The areas that affect the probability the most are the most important in the image.

=) Third approach is to calculate the gradient of a neuron with respect to each input pixel. This will us an image of gradients. Each image can be seen and wherever the value of gradient is high, more whiter pixels will appear in the output of this step with respect to that neuron. To calculate these gradients we can use back propogation. One problem with this method is that gradients can be highly positive or negative in areas of importance, but when visualised through a digital image, these gradients cancel the effect of each other in these areas and that leads to blunt results and noisy output.

To deal with this problem, we use guided backpropogation.
=) In this method, we choose a neuron of interest and set all other neurons in that layer to 0. Then we input an image to the network and perform a feed forward step. Then we perform the backpropogation step, to calculate the gradient matrix of this neuron w.r.t. to the input. But instead of allowing negative gradient to flow back, we clip negative gradients to 0. This is done to all neurons in all the layers. This prevents the effect discussed earlier and the results become more vivid in the resultant image. Basically we focus only on the positive influencies (gradients)

=) One more method to interpret the results of a CNN is to optimize over images. The goal here is to create an image which gets classified as a certain class. So, if this image is passed through a trained ConvNet it should maximize the probability of the class of interest. Doing this would allow us to visualize what images are classified in which manner.

To do this we can set this as an optimization problem with respect to the pixels of the input image.

$$\text{argmax}\left(\text{Score}_c(I) - \lambda \mathcal{L}(I)\right) = \text{argmax}\left(-L(\theta)\right)$$

where $\text{Score}_c(I)$ = Score for class before softmax

$\mathcal{L}(I)$ = some regularizer to ensure that I looks like an image.

In the whole process we keep the parameters of the convolutional neural network fixed. We adjust the image pixels so that the score of the class is maximized.

⇒ We start with a zero image / randomly initialized image.

⇒ Set the score vector to be $[0\ 0\ \ldots 1 \ldots 0\ 0]$

⇒ Do a forward pass.

⇒ Compute the gradients $\frac{\partial L(\theta)}{\partial I[i,j]}$ for all i,j

⇒ Update the pixel $I[i,j] = I[i,j] - \eta \nabla_{I[i,j]} L(\theta)$

⇒ Repeat the process from forward pass.

Instead of limiting this process to only the output, we can focus on any one neuron in any layer to see what images makes that neuron to fire.

To do this

⇒ Feed an image through the network. (The propagation can stop at the neuron of interest)

⇒ Set activation in layer of interest to all zero, except for the neuron of interest.

⇒ Backprop to image

⇒ $I[i,j] = I[i,j] - \eta \nabla_{I[i,j]} L(\theta)$, where $L(\theta)$ can be negative of the pre-activation of the neuron of interest.

Another similar idea is to create images using embeddings.

So there is a pre-trained CNN, now the objective is to feed this network with an image, not the embeddings of each layer, and then create/reconstruct an image such that the created image's embeddings are as close to the original image's embeddings.

Embeddings are the output of each layer.

So here the loss function is the squared error between the embeddings of the orignal image and the created image.

$$L(I) = \| \phi(I) - \phi(I_o) \|^2 + \lambda \| \phi(x) \|_6^6$$

$\phi(I)$ = vector of embeddings of the created image

$\phi(I_o)$ = vector of embeddings of the original image.

$$I[i,j] = I[i,j] - \eta \frac{L(I)}{I[i,j]}$$

=) This can be done for embeddings obtained from any layer.

=) We observe that the created image gets more abstract as we take embeddings from deeper layers.

Deep Dream

Suppose instead of starting with a blank image, we start with an actual image. Now we select some neuron of the network and our objective is to change the image such that this neurons fire even more.

our objective becomes $\max L(I)$

where $L(I) = h_{ij}^2$

Where h_{ij} is the output of the j^{th} neuron of the i^{th} layer.

We can repeat the same procedure of forward propogation and back propogation and update the image pixels $I[i,j] = I[i,j] + \eta \frac{\nabla L(I)}{I[i,j]}$

Doing this iteratively would make the image more and more like the patterns that cause the neuron to fire.
The Network has been trained to detect certain patterns which appear frequently in the training dataset. So it starts to see those patterns even when they hardly exist.

=) For example, if a cloud looks a little bit like a bird, the network will make it look more like a bird. This in turn will make the network recognize the bird even more strongly on the next pass, until a detailed bird appears in the image.

Deep Art

The objective of this is to ~~input~~ provide two images; a Content image, and a Style image, and then create an image such that the content of the created image is same as the content image but it is in the style of the style image.
This is done through matching the embeddings of ~~the~~ both the content image and Style image with the created image.

For Capturing the content of the image:-

$$L_{content}(\vec{p}, \vec{x}) = \sum_{ijk} (p_{ijk} - x_{ijk})^2$$ where $\vec{p}$ is the embeddings of the content image. $\vec{x}$ is the embeddings of the created image. This is done for some layer and the formula is written in the tensor notation where the output embeddings is in the form of tensor.

For capturing the style of the image:-
It turns out that if $V \in \mathbb{R}^{64 \times (256 \times 256)}$ is the activation/output at a layer, then, $V^T V \in \mathbb{R}^{64 \times 64}$ captures the style of the image.

The deeper layers capture more of this style information.

The objective can be the the following:-

$L_{style} = E_l = \sum_{ij} (G^l_{ij} - A^l_{ij})^2$, where G^l is the Gram matrix of the style image computed at layer L, and A^l is similarly for the created image.

⇒ The total loss is given as:- $L_{total}(I) = \alpha L_{content}(I) + \beta L_{style}(I)$

Fooling Deep CNNs

By optimizing over images, we can also fool deep CNN's to wrongly predict classes for obvious images. This is done through maximizing the log likelihood of the incorrect class and making changes in the images. The resultant images fools the network into predicting the wrong classes.

Now if instead, we gave blank image as input, then the resultant image is very noisy and un-interpretable; but, even then the network convincingly predicts it as belonging to a class.

This happens because images are very high dimensional entities. Out of these high dimensional images, only a few are interpretable by the human brain. But for a CNN, each image in this space is just a vector, and it just passes that vector through a learnt function which has its decision boundary. Even though the image is noisy to us, the CNN predicts it according to the output of the learnt function.

Deep Learning Part-2

Representing Joint Distributions:- Suppose there are 3 random variables :- I takes 2 values, G takes 3 values, S takes 2 values. Now if we want to represent their joint distribution, we would require $2 \times 3 \times 2 - 1 = 11$ values, last one is 1 - (sum of 11). If we use chain rule for representing the joint representation :- $P(S,G,I) = P(S|G,I) \times P(G|I) \cdot P(I)$

$P(I)$ can be represented with 1 value :- $P(I_1)$ or $P(I_2)$

$P(G|I)$ can be represented with 4 values :- $P(G_1|I_1), P(G_2,I_1)$ $P(G_1|I_2), P(G_2,I_2)$

$P(S|G,I)$ can be represented with 6 values.

$P(S_1|$ 6 different combinations of I and G)

$P(G_3|I_1)$ and $P(G_3|I_2)$ can be calculated/derived

In total we still take $1+4+6 = 11$ values to represent the joint distribution. But if we make some assumptions on the conditional independence of S and G given I, then the number of parameters can be reduced. So $P(S|G,I) = P(S|I)$. Hence the joint $P(G|S,I) = P(G|I)$ distribution can now be written as :- $P(S,G,I) = P(S|I) \cdot P(G|I) \times P(I)$

2 values $P(S_1|I_1)$ and $P(S_1|I_2)$ — 4 values — 1 value.

The number of parameters reduces to $2+4+1 = 7$ values.

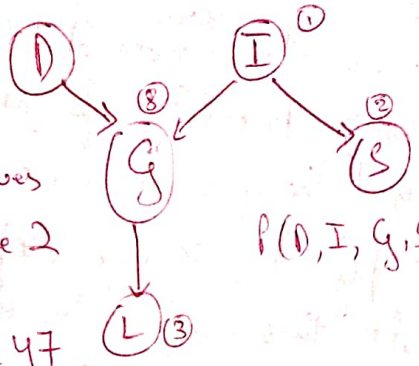
Assumptions like this can be made based on the application, the advantages of this representation are :- 1) more natural, variables that exhibit dependence on others is made explicit.

2) More compact, less parameters required if conditional independence is assumed.

3) More modular, these values can be reused if more variables are added later, reusing values is not an option in plain representation.

Bayesian Networks :- Nodes are random variables and directed edges specify the dependence relations between the random variables.

For example :-



The joint distribution can be specified as:-

$P(D, I, G, S, L) = P(D) \cdot P(I) \cdot P(S|I) \cdot P(G|D,I) \cdot P(L|G)$

If, G takes 3 values and all others take 2 values, the plain representation takes = 47 values.

will the compact representation takes $1+1+2+8+3 = 15$ values.

The graph structure along with conditional probability distributions is called a Bayesian Network.

⇒ The independence assumptions are made by us as a modelling choice and the number of parameters and data required to train the model is a consequence of our modelling choice.

Reasoning Encoded in Bayesian Network :-

i) **Causal Reasoning**:- we try to predict downstream effects of various factors. For example asking question like what is the chance, the student gets a good LOR (L^1) given that they are intelligent (I^1). Intelligence (I) is a factor and how it affects L is the question that is being asked here. $P(L^1|I^1)$

$$P(L^1|I^1) = \frac{P(L^1, I^1)}{P(I^1)}$$

, we expect that $P(L^1|I^1) > P(L^1)$ since the chance of getting a good LOR increases given the student is intelligent.

ii) **Evidential Reasoning**:- We reason about causes by looking at their effects. Asking questions like what is the probability of the student being intelligent given that he got a poor Grade(G^3), $P(I^1|G^3)$, we expect $P(I^1|G^3) < P(I^1)$, same for difficulty $P(D^1|G^3) > P(D^1)$, the difficulty of course is most probably hard given the student scored poorly. (increases)

(ii) Explaining Away :- We see how different causes of the same effect can interact. We already saw that $P(I'|G^3) < P(I')$ But what if we are given another cause of the grade, the difficulty. So probability of student being intelligent given he scored poorly and the course was difficult. $P(I'|G^3, D') > P(I'|G^3)$, is our expectation since now the course was also difficult, this compensates the drop in intelligence. So in a way, various causes explain the variation in other causes.

Independencies Encoded by a Bayesian Network :-

Given n random variables, we are interested in knowing if $X_i \perp X_j$ or $X_i \perp X_j \mid Z$ where $Z \subseteq X_1, X_2 \cdots, X_n / X_i, X_j$

Rule 1 :- A node is not independent of its parents. (By construction)

Rule 2 :- A node is not independent of its parents even when we are given the values of other variables. For example knowing the value of I does not mean that L and G become independent. L still depends on G even after knowing I.

These were some independencies between nodes and their parents. Now what about non-parents. L can still be affected by S which is a non-parent, since S affects I and I affects G which then affects L. So $L \not\perp S$, but given the value of G, $L \perp S$ holds. So a node is independent of non-descendents given the parents. So $L \perp S \mid G$, moreover $L \perp I \mid G$ also holds. This is because, the effect S or I have on L is through G, but we already know G, so knowing S or I does not matter for L.

Independencies of a node and its descendants is also encoded.

Rule 3 :- A node is independent to all non-descendents given the parent variables of the node.

Formally :- A Bayesian Network is a directed acyclic graph where nodes represent variables. Let P_{X_i} denote the parents of X_i

and $ND(X_i)$ denotes the non-descendants of X_i (does not include parents). Then for each variable X_i the graph encodes the following conditional independence assumptions:-

$$I_i(G) = \left((X_i \perp ND(X_i)) \mid P_{X_i} \right)$$

I-Maps :- Let P be a joint distribution over $X = X_1, X_2, \cdots, X_n$. $I(P)$ be the set of independence assumptions that hold in P. Each element of this set is of the form $X_i \perp X_j \mid Z$, $Z \subseteq X \setminus \{X_i, X_j\}$. Let $I(G)$ be the set of independence assumptions associated with a graph G. We say that G is an I-map for P if $I(G) \subseteq I(P)$. Hence all the assumptions that G states hold in P. But P can have additional independencies.

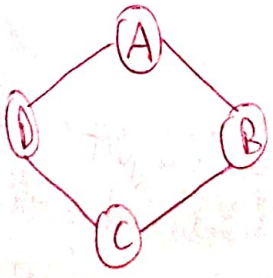
In practice, we do not know P and hence can't compute $I(P)$. So we just make assumptions about $I(P)$ and then construct a G such that $I(G) \subseteq I(P)$.

Theorem :- Let G be a BN over a set of random variables X and let P be a joint distribution over these variables. If G is an I-Map for P, then P factorizes according to G.

Theorem :- Let G be a BN over a set of random variables X and let P be a joint distribution over these variables. If P factorizes according to G, then G is an I-Map of P.

Perfect Map :- A graph G is a Perfect Map for a distribution P if the independence relations implied by the graph are exactly the same as those implied by the distribution.

Undirected Graphical Models :-

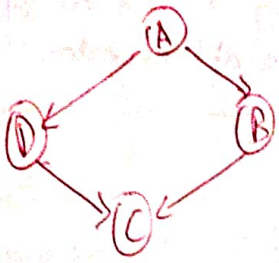


The conditional independencies in this problem are:-

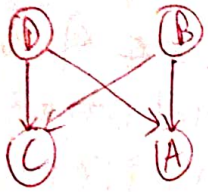
$A \perp C \mid \{B, D\}$ (Because A and C never interact)

$B \perp D \mid \{A, C\}$ (Because B and D never interact)

What if we try to model this using Bayesian Networks?



this captures $A \perp C \mid \{B, D\}$ but it does not capture $B \perp D \mid \{A, C\}$ because given C and B can affect D (Explaining Away)

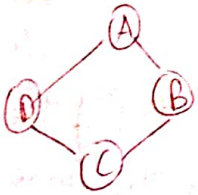


$A \perp C \mid \{B, D\}$ is encoded
But it also encodes
$B \perp D$ (unconditional)

> Turns out that no Bayesian Network can exactly capture independence relations completely (Perfect Map)

This is because a directed edge between two nodes implies some kind of direction in the interaction. But in our example A and B are equal partners. We instead want to capture the strength of the interaction between the nodes. These undirected models are called Markov Networks.

We can't use Conditional Probability Distributions as they don't make sense in the undirected case since there is no natural conditioning. Hence the parameters in this case is the affinity strength between connected random variables. We can have factors $\phi_1(A,B)$, $\phi_2(B,C)$, $\phi_3(C,D)$, $\phi_4(A,D)$ which capture the affinity between the corresponding nodes. Suppose All variables take 2 values, then:-



$\phi_1(A,B)$			$\phi_2(B,C)$	...	$\phi_4(A,D)$	
a^0	b^0	30	b^0	c^0	a^0	d^0
a^0	b^1	5				
a^1	b^0	1				
a^1	b^1	10				

Notice that here parameters are defined over the edges, while in BN's the parameters were defined over the nodes.

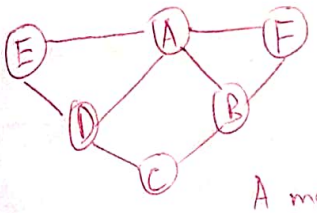
These tables are not probability distributions. These are just weights which can be interpreted as the relative likelihood of an event. To convert these into probability distributions we can normalize the product :-

$$P(a,b,c,d) = \frac{1}{Z} \times \phi_1(a,b) \cdot \phi_2(b,c) \cdot \phi_3(c,d) \cdot \phi_4(d,a)$$

where $Z = \sum_{a,b,c,d} \phi_1(a,b) \cdot \phi_2(b,c) \cdot \phi_3(c,d) \cdot \phi_4(a,d)$

$\rightarrow$ for all values of a, b, c, d.

Combinations.



We could still factorize pairwise or we could factorize based on maximal cliques.

A maximal cliques is a set of nodes in which adding any other will result in a non-clique set.

A clique is a set of nodes which are pairwise connected to each other.

So the above network can be factorized like this:-

$\phi_1(A,D,E) \cdot \phi_2(A,B,F) \cdot \phi_3(B,C) \cdot \phi_4(C,D)$

So a distribution factorizes over a Markov Network H if P can expressed as $P(X_1, \cdots, X_n) = \prod_{i=1}^{m} \phi_i(D_i)$ where D_i is a complete sub graph (clique in H)

A distribution is a Gibbs distribution parametrized by a set of factors $\Phi = \{\phi_1(D_1), \cdots, \phi_m(D_m)\}$ if it is defined as $P(X_1, X_2, \cdots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(D_i)$

=) Let X, Y, Z be some distinct subsets of U (set of all Random variables), then a distribution P over these RVs would imply $X \perp Y \mid Z$ if and only if we can write $P(X) = \phi_1(X,Z)\ \phi_2(Y,Z)$

In our example of 4 variables. $P(a,b,c,d) = \frac{1}{Z} \phi_1(a,b) \cdot \phi_2(b,c) \cdot \phi_3(c,d) \cdot \phi_4($

can be written as :- $P(a,b,c,d) = \frac{1}{Z} \phi_5(b,\{a,c\}) \cdot \phi_6(d,\{a,c\}$

Hence $B \perp D \mid \{A,C\}$ holds. Similarly $A \perp C \mid \{B,D\}$ also hold

For a given Markov Network H we define Markov Blanket of a RV X to be the neighbor of X in H.

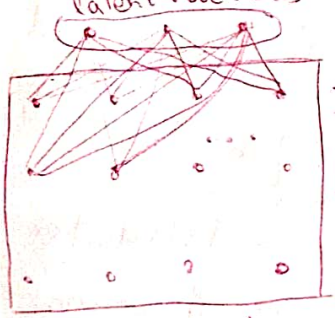
Analogous to the case of Bayesian Networks we can define the local independencies associated with H to be :-

$$X \perp (U - \{X\} - MB_H(X)) \mid MB_H(X)$$

X is independent of its non-neighbors given the neighbors of X in H. The neighborhood is defined by the Markov Blanket. For example in an image, the Markov Blanket can be the 8 pixels that surround it.

Latent Variables:- Now in the case of modelling an image as a Markov Network, each pixel can be treated as a random variable with blanket being the 8 surrounding pixels. So the joint distribution of images can be $P(V_1, V_2, \cdots V_{1024})$ where V_i is a pixel and can be factorized accordingly.

Instead of just modelling the pixels as random variables, we can introduce latent variables such that the value of pixels depend on the ~~visible~~ latent variables. Consider the task of generating a scenery. The latent variables can be:- is the scene sunny or not, does the scene have mountains or beaches, clouds or not etc. Based on these variables, the actual pixels can be generated. The Markov Network looks like this:-



pixels are dependent on the latent variables. The interactions between the pixels are captured through the latent variables.

This modelling choice helps in many tasks like abstraction. Suppose, we are able to learn the joint distribution $P(V,H)$. Using this distribution we can find

$$P(H|V) = \frac{P(V,H)}{\sum_V P(V,H)}$$ → marginalize over V.

So given an image, we can find the most likely latent configuration of H that generated this image. The vector h calculated can be thought of as an abstract representation of the whole image. Similar images which are pixel wise still very different in the original space are expected to have very similar latent configuration.

⇒ In practice we don't have latent configurations given to us. This is like a Representation Learning Task and each element in the latent vector representation may not even be interpretable they just help us to learn a good abstraction of the data.

⇒ Another task these representations can help is generation. Using the joint distribution, we can find :- $P(V|H) = \frac{P(V,H)}{\sum_V P(V,H)}$

So, given a latent configuration, generate the image which most likely corresponds to this configuration.

Let the vector V of all visible variables be a binary vector, $V \in \{0,1\}^m$ and the vetor H also be a binary vector $H \in \{0,1\}^n$.

Notice that edges are only incident between each pair of hidden, -visible variables. We donot have edges between (hidden, hidden) and (visible - visible) variables. Hence the cliques in the graph only includes each pair of (hidden, visible) pair of variables. So the joint distribution can be factorized as :- $P(V,H) = \frac{1}{Z} \prod_{i=1}^{m} \prod_{j=1}^{n} \phi_{ij}(v_i, h_j)$

We can add additional factors such that

$$P(V,H) = \frac{1}{Z} \prod_{i=1}^{m} \prod_{j=1}^{n} \phi_{ij}(v_i,h_j) \prod_{i=1}^{m} \psi_i(v_i) \prod_{j=1}^{n} \varepsilon_j(h_j)$$

this is allowed as long as Z correctly normalizes to obtain a prob distribution. Z is calculated as follows :-

$$Z = \sum_V \sum_H \prod_{i=1}^{m} \prod_{j=1}^{n} \phi_{ij}(v_i,h_j) \prod_{i=1}^{m} \psi_i(v_i) \prod_{j=1}^{n} \varepsilon_j(h_j)$$

all 2^m possible configurations → all 2^n possible configurations. Hence this Z term includes 2^{m+n} terms in total.

Each ϕ_{ij} is a table which gives the values for each possible (4) configuration of v_i and h_j. Similarly ψ_i and ε_j are also tables giving values for each possible (2) configuration of v_i and h_j respectively.

Restricted Boltzmann Machines:- TO learn ϕ_{ij}s, and ψ_is and ε_js, we introduce parameters. RBMs chooses the following parametric forms:- $\phi_{ij}(v_i, h_j) = e^{w_{ij} v_i h_j}$

With this parametric form, the distribution can be expressed as follows:-

$$\psi_i(v_i) = e^{b_i v_i}$$

$$\varepsilon_j(h_j) = e^{c_j h_j}$$

$$P(V,H) = \frac{1}{Z} \prod_{i=1}^{m} \prod_{j=1}^{n} e^{w_{ij} v_i h_j} \prod_{i=1}^{m} e^{b_i v_i} \prod_{j=1}^{n} e^{c_j h_j}$$

$$= \frac{1}{Z} e^{\left(\sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} v_i h_j + \sum_{i=1}^{m} b_i v_i + \sum_{j=1}^{n} c_j h_j \right)}$$

$$\boxed{P(V,H) = \frac{1}{Z} e^{-E(V,H)}} \quad \text{where} \quad \boxed{E(V,H) = -\sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} v_i h_j - \sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} c_j h_j}$$

$E(V,H)$ is called the energy

High energy → low probability, and vice versa.

This is called RBM because, we donot allow edges between the visible vertices (restricted), and the form of joint distribution is similar to Boltzmann/Gibbs distribution.

RBMs as stochastic Neural Networks:- Let us try to model $P(H|V)$ and $P(V|H)$ given the form of joint distribution.

Let $V_{-\ell}$ be the state of all visible units except the ℓ^{th} unit.

$$\alpha_\ell(H) = -\sum_{j=1}^{n} w_{\ell j} h_j - b_\ell \quad \Big] \text{ part of } \ell^{th} \text{ visible unit in the summation of } E(V,H).$$

$$\beta(V_{-\ell}, H) = -\sum_{\substack{i=1 \\ i \neq \ell}}^{m} \sum_{j=1}^{n} w_{ij} v_i h_j - \sum_{\substack{i=1 \\ i \neq \ell}}^{m} b_i v_i - \sum_{j=1}^{n} c_j h_j$$

So $E(V,H) = v_\ell \alpha_\ell(H) + \beta(V_{-\ell}, H)$

We can hence calculate $P(V_\ell = 1 \mid H)$. Notice that since $V_{-\ell}$ is independent of V_ℓ given H. $P(V_\ell = 1 | H) = P(V_\ell = 1 | V_{-\ell}, H)$ doesn't matter what configuration $V_{-\ell}$ is.

Representing it in this way allows for using the joint distribution :- $P(V_\ell = 1 \mid H) = P(V_\ell = 1 \mid V_{-\ell}, H) = \frac{P(V_\ell, V_{-\ell}, H)}{P(V_{-\ell}, H)}$

$$\Rightarrow \frac{e^{-E(V_\ell = 1, V_{-\ell}, H)}}{e^{-E(V_\ell = 0, V_{-\ell}, H)} + e^{-E(V_\ell = 1, V_{-\ell}, H)}} \Rightarrow \frac{e^{-\alpha_\ell(H) - \beta(V_{-\ell}, H)}}{e^{-\alpha_\ell(H) - \beta(V_{-\ell}, H)} + e^{-\beta(V_{-\ell}, H)}}$$

$$\Rightarrow \frac{e^{-\beta(V_{-\ell}, H)} \cdot e^{-\alpha_\ell(H)}}{e^{-\beta(V_{-\ell}, H)} \cdot e^{-\alpha_\ell(H)} + e^{-\beta(V_{-\ell}, H)}} \Rightarrow \frac{e^{-\alpha_\ell(H)}}{e^{-\alpha_\ell(H)} + 1} = \frac{1}{e^{\alpha_\ell(H)} + 1}$$

$$\Rightarrow \boxed{P(V_\ell = 1 \mid H) = \sigma(-\alpha_\ell(H)) = \sigma\left(\sum_{j=1}^{n} w_{\ell j} h_j + b_\ell\right)}$$

Similarly it can be shown that $\boxed{P(h_p = 1 \mid V) = \sigma\left(\sum_{i=1}^{m} w_{ip} v_i + c_p\right)}$

Hence, the extra factors added to the form of $P(H,V)$ serve as biases in the formulation of RBMs. So RBM's can be interpreted as stochastic neural network where the nodes and edges correspond to neurons and synaptic connections, respectively. The conditional probability of a single variable being 1 can be interpreted as the firing of a neuron with sigmoid activation.

Having $P(v_\ell = 1 \mid H)$, the Probability $P(V \mid H)$ is straightforward.

$P(V \mid H) = \prod_{i=1}^{m} P(V_i = v_i \mid H)$ → since V_i are conditionally independent independent.

$$\rightarrow \begin{cases} \sigma(-\alpha_\ell(H)) & \text{if } v_i = 1 \\ 1 - \sigma(-\alpha_\ell(H)) & \text{if } v_i = 0 \end{cases}$$

Since our task is unsupervised, our objective function should only depend on V's → the images. Eventually we want to maximize the likelihood of our data. So the task is now to find parameters w_{ij}, b_i, c_j that maximize the probability of the images. So we want to maximize $P(V \mid \theta)$ → the set of params

So maximize $\prod_{i=1}^{m} P(V_i = v_i | \theta)$. Lets consider the loss for one single example. $P(V_i = v_i)$ is same as $P(V_i = v_i | \theta)$ (this just means based on the current params.)

So $P(V) = \prod_{i=1}^{m} P(V_i = v_i)$

Also we know that $P(V,H) = \frac{1}{Z} e^{-E(V,H)}$. So $P(V)$ can also be calculated $\Rightarrow \frac{1}{Z} \sum_{H} e^{-E(V,H)}$

$\sum_H$ → marginalized over H. all possible 2^n configuration.

So $P(V) = \frac{1}{Z} \sum_H e^{-E(V,H)} = \frac{\sum_H e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}}$

Z, marginalized over 2^{n+m} configuration of V and H.

So $\ln P(V) = \ln \sum_H e^{-E(V,H)} - \ln \sum_{V,H} e^{-E(V,H)}$

(we maximize this instead)

$$\frac{\partial \ln P(V)}{\partial \theta} \rightarrow \text{any param} = \frac{-1}{\sum_H e^{-E(V,H)}} \times \sum_H e^{-E(V,H)} \cdot \frac{\partial E(V,H)}{\partial \theta} + \frac{1}{\sum_{V,H} e^{-E(V,H)}} \times \sum_{V,H} e^{-E(H,V)} \cdot \frac{\partial E(H,V)}{\partial \theta}$$

$$= -\sum_H \frac{e^{-E(V,H)}}{\sum_H e^{-E(V,H)}} \frac{\partial}{\partial \theta} E(V,H) + \sum_{H,V} \frac{e^{-E(H,V)}}{\sum_{V,H} e^{-E(H,V)}} \cdot \frac{\partial}{\partial \theta} E(H,V)$$

(after dividing both numerator and denom by Z, observe that)

$\frac{P(V,H)}{P(V)} = P(H|V)$ $\qquad$ $P(V,H)$

Hence $$\boxed{\frac{\partial \ln P(V)}{\partial \theta} = -\sum_H P(H|V) \cdot \frac{\partial E(V,H)}{\partial \theta} + \sum_{H,V} P(V,H) \cdot \frac{\partial E(H,V)}{\partial \theta}}$$

if $\theta = w_{ij}$ then

$$\frac{\partial L(\theta)}{\partial w_{ij}} = + \sum_{H} p(H|V) h_j v_i - \sum_{H,V} p(V,H) h_j v_i$$

$$= + E_{p(H|V)}[h_j v_i] - E_{p(H,V)}[h_j v_i]$$

Both these expectations have an exponential number of terms. And notice that this is gradient wrt 1 param of a single observation. Hence the learning is intractable in its current form. So we need sampling methods to estimate these expectations. But sampling is not straightforward as only a few samples are likely from the huge space of samples. So sampling has to be on the basis of the joint distribution but we don't have the distribution itself.

Solution is to draw samples from an easier distribution as long as sampling from the other distribution is same as if they were drawn from P(original). We do this in Gibbs sampling.

Goal 1: Given a random variable $X \in \mathbb{R}^n$ we are interested in drawing samples from the joint distribution $P(X)$

Goal 2: Given a function $f(X)$ defined over the random variable X, we are interested in computing the expectation $E_{P(X)}[f(X)]$ (useful for gradient calc in RBMs)

Suppose we have a chain of random variables $X_1, X_2, \ldots X_k$ each $X_i \in \mathbb{R}^n$, basically each Random variable X_i is in itself a vector. i corresponds to a time step. For our discussion, let $X_i \in \{0,1\}^n$. One way of looking is to see that X_i transitions to X_{i+1} when the state changes from i to $i+1$. X_i can take one of 2^n values in each state. One might be interested in finding the most likely value X_i will taken given all the values of previous states $X_{i-1}, X_{i-2}, \ldots X_1$.

So we want :- $P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots X_{i-1} = x_{i-1})$

But let the chain exhibit the Markov Property:-

$$P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots X_{i-1} = x_{i-1}) = P(X_i = x_i | X_{i-1} = x_{i-1})$$

given the previous state X_{i-1}, X_i is independent of all preceding states. $X_1 \to X_2 \to \cdots \to X_k$ $\left[X_i \perp\!\!\!\perp X_1^{i-2} \mid X_{i-1}\right]$.

This chain of random variables is called a Markov Chain. We are dealing with a discrete time and discrete space Markov Chain.

(i is discrete; X_i takes discrete values.)

Let $2^n = l$. Now to specify $P(X_i = x_i \mid X_{i-1} = x_{i-1})$ we need to specify l^2 values, for each configuration of X_{i-1} and X_i. We can represent this as a matrix $T \in l \times l$ where T_{ab} gives the probability of transitioning from $X_{i-1} = a$ to $X_i = b$., where 'a' and 'b' encode two of the possible 2^n configurations respectively. T is called the transition matrix. T may be different for each transition, but we make the assumption that our Markov Chain is time homogeneous. So for all transitions, T remains the same. $\boxed{P(X_i = b \mid X_{i-1} = a) = T_{ab}} \quad \forall a, b, \forall i$

Let the distribution be specified by μ^k at k^{th} time step. μ^k is a l dimensional vector. $\mu^k_a = P(X_k = a)$. The a^{th} element of μ^k gives the probability of X_k being 'a'. Again, X_k is an n-dimensional vector but 'a' is an encoding of one of such possible vector.

('a' is an integer)

Let the starting distribution be μ^0. How to calculate $P(X_1 = b)$ from μ^0?

$$P(X_1 = b) = \sum_a P(X_0 = a, X_1 = b) \quad \text{marginalize over } X_0.$$

($P(X_1 = b)$: Probability of X_1 taking b irrespective of the starting point)

$$= \sum_a P(X_0 = a) \cdot P(X_1 = b \mid X_0 = a)$$

$$= \sum_a \mu^0_a T_{ab} = \mu'_b \quad \text{since } P(X_1 = b) = \mu'_b$$

So μ' can be written as $\mu' = \mu^0 T = [\mu^0_1 \ \mu^0_2 \cdots \mu^0_l] \begin{bmatrix} T_{11} & & \\ T_{21} & & \\ \vdots & & \\ T_{l1} & & \end{bmatrix}$

Similarly $\mu^2 = \mu^1 T^2$

So in general $\boxed{\mu^k = \mu^0 T^k}$ → this gives us a way to get the distribution at time k. To get the random variables at k, we can perform sampling from μ^k.

This is still computationally expensive as the dimensions of the vectors involved are exponential in terms of the number of variables.

Also notice that if at some time step t, u^t reaches a distribution π such that $u^t T = u^t = \pi$. Then for all subsequent time steps $u^j = \pi \quad \forall j \geq t$

π is called the stationary distribution of the Markov Chain

$X_t, X_{t+1}, X_{t+2}, \cdots$ will all follow the same distribution π.

So if we run a Markov Chain for a large number of time steps then after a point we start getting samples $x_t, x_{t+1}, x_{t+2}, \cdots$ which are essentially being drawn from the stationary distribution. But as mentioned earlier, drawing from this distribution is computationally expensive and still intractable.

But suppose if we are given a Markov Chain such that it is easy to draw samples from it and the stationary distribution of this chain is $P(X)$, Then we could empirically estimate $E_{P(x)}[f(x)]$ as $\frac{1}{n} \sum_{i=t}^{t+n} f(X_i)$

→ the time step after we get the stationary distribution.

Theorem:-

If $X_0, X_1, \cdots X_t$ is an irreducible time homogeneous discrete Markov Chain with stationary distribution π, then

$\frac{1}{t} \sum_{i=1}^{t} f(X_i)$ converges almost surely $\xrightarrow{t \to \infty} E_\pi[f(X)]$ where,

for any function $f: \mathcal{X} \to \mathbb{R}$

$X \in \mathcal{X} \to \{0,1\}^n$ in our case

$X \sim \pi$

If further the Markov Chain is aperiodic then

$$P(X_t = x_t \mid X_0 = x_0) \to \pi(X) \quad \text{as } t \to \infty \quad \forall x, x_0 \in \mathcal{X}$$

$$T = \begin{bmatrix} T_{11} \\ T_{21} \\ T_{t1} \end{bmatrix}$$

Basically means that doesn't matter the starting state, after large number of steps, the samples calculated are drawn from $\pi(X)$

Now our task is to find such a Markov Chain whose $\pi(X)$ is same as $P(X)$, moreover it is irreducible and aperiodic.

For ease of notation, instead of $X = V_1, V_2, \ldots V_m, H_1, H_2, \ldots, H_n$ We will use $X = X_1, X_2, \ldots X_{n+m}$ (Basically renaming them all to X).

We will now set up a Markov Chains for RBMs.

Here X_i does not mean random vector at time step i, here it means the i^{th} element of the random vector, basically a scalar 0 or 1.

We set up the chain in this way:- i) We start, at a random X where each element is set to 0 or 1 uniformly at random.

ii) Sample a value $i \in \{1, \text{to } n+m\}$ (index) using a distribution, $q(i)$, can be uniform.

iii) Fix the value of all variables except X_i. Sample a new value for X_i using the following conditional distribution

$P(X_i = y_i | X_{-i} = x_{-i})$

↳ conditioned on remaining variables.

So a transition from state x to state y is only possible if x and y differs at most by 1 variable out of $n+m$ variables. All other transitions are not possible in this chain. Hence the transition matrix is extremely sparse and can be explained like this:-

$$T_{xy} = \begin{cases} q(i)\, P(X_i = y_i | X_{-i} = x_{-i}), & \text{if } \exists\, i \in \{1 \text{ to } n+m\} \\ & \text{such that if } v \neq i \quad x_v = y_v \\ 0, \text{ otherwise} \end{cases}$$

Observe that in the case of RBMs, $P(X_i = y_i | X_{-i} = x_{-i})$ is easily calculated. If $i \leq m$ (visible variable is selected)

$$P(V_i = 1 | V_{-i}, H) = P(V_i = 1 | H) = \sigma\left(\sum_{j=1}^{n} w_{ij} h_j + b_i\right)$$

and if $i > m$, (hidden variable is selected)

$$P(H_j = 1 | V, H_{-i}) = P(H_i = 1 | V) = \sigma\left(\sum_{i=1}^{m} w_{ij} v_i + c_j\right)$$

Hence the sampling procedure at each step is as follows:—

=) Sample i from $q(i)$

=) Sample X_i from a Bernoulli Distribution, keeping all variables the same. Both these computations are easy. Hence it is easy to sample from the chain.

Theorem (Detailed Balanced Condition):— To show that a distribution π is a stationary distribution for a Markov Chain described by the transition probabilities T_{xy}, it is sufficient to show that $\forall\, x, y$ the following condition holds:—

$$\boxed{\pi(x)\, T_{xy} = \pi(y)\, T_{yx}}$$

. So in our case if we want to prove that $P(X)$ is actually the stationary distribution of this chain then it suffices to show that $P(x)\, T_{xy} = P(y)\, T_{yx}$

To prove this, we look at 3 cases, when x and y differ in more than variables, 1 variable, are same. i) differ in more than 2 variables:— In this case $T_{xy} = 0$ and $T_{yx} = 0$, hence $P(x)\, T_{xy} = P(y)\, T_{yx} = 0$ holds true

ii) when x any y are same so $P(x) = P(y)$ and $T_{xy} = T_{yx} = T_{xx}$

So $P(x)\, T_{xy} = P(x)\, T_{xx} = P(y)\, T_{yx}$, holds true.

iii) when x and y differ in 1 variable then, let that variable be the i^{th} variable.

So, $T_{xy} = q(i)\, P(X_i = y_i \mid X_{-i} = x_{-i})$

$$P(x)\, T_{xy} = P(x_i, x_{-i}) \cdot q(i) \cdot \frac{P(y_i, x_{-i})}{P(x_{-i})}$$

$$= P(y_i, x_{-i}) \cdot q(i) \cdot \frac{P(x_i, x_{-i})}{P(x_{-i})}$$

(Since other variables are same)

$$= P(y) \cdot q(i) \cdot P(X_i = x_i \mid X_{-i} = x_{-i})$$

$$P(x)\, T_{xy} = P(y) \cdot T_{yx} \qquad \left[\because T_{yx} = P(X_i = x_i \mid X_{-i} = y_{-i})\right]$$

Hence the detailed balance condition holds and $P(X)$ is indeed the Stationary Distribution of this chain.

Now we need to prove that this chain is irreducible and aperiodic.
A markov chain is irreducible if one can get from any state in the sample space to any other state in a finite number of transitions.
$\forall x,y \quad \exists k>0$ with $P(X^{(k)}=y \mid X^{(0)}=x) > 0$.

Our Markov chain is irreducible as from any starting state, any other state only differs in finite number of variables, and those variables can be flipped in one by one with a non zero probability.

A chain is called aperiodic if $\forall i$ in the sample space, the greatest common divisor of the set :-
$\{k \mid P(X^{(k)}=x \mid X^{(0)}=x) > 0)$ is 1. This set is the set of all those time steps such that the initial value is obtained again. Having a GCF of 1 essentially means that these numbers are arbitrary and are not a multiple of any number, which would otherwise make the chain periodic. This has to satisfy for all x in the sample space. Our Markov chain is aperiodic since the initial value can be retained in each time step with a non-zero prob, and gcf of all natural numbers is 1.

We return to gradient computation of RBMs.

$$\frac{\partial L(\theta)}{\partial w_{ij}} = \underset{P(H|V)}{E}[h_j v_i] - \underset{P(H,V)}{E}[v_i h_j] = \sum_H p(h|v) h_j v_i - \sum_{V,H} p(v,h) h_j v_i$$

$$= \sum_H p(h|v) h_j v_i - \sum_V p(v) \sum_H p(H|v) h_j v_i$$

$$\sum_H p(h|v) h_j v_i = \sum_{H_j} \sum_{H_{-j}} p(h_j|v)\, p(h_{-j}|v)\, h_j v_i$$

H_j: 2 values; H_{-j}: 2^{n-1} values here.

$$= \sum_{H_j} p(h_j|v) h_j v_i \underbrace{\sum_{H_{-j}} p(h_{-j}|v)}_{\text{this just sum of a distribution. Hence equal to 1.}}$$

$$= p(h_j=1|v) v_i + p(h_j=0|v)\, v_i \times 0$$

$$= p(h_j=1|v) v_i = \sigma\left(\sum_{i=1}^{m} w_{ij} v_i + c_j\right) v_i$$

Hence $\frac{\partial L(\theta)}{\partial w_{ij}} = \sigma\left(\sum_{i=1}^{m} w_{ij} v_i + c_j\right) v_i - \sum_V p(v)\, \sigma\left(\sum_{i=1}^{m} w_{ij} v_i + c_j\right) v_i$

we can write this expression as a general for the whole matrix W.

$$\boxed{\frac{\partial L(\theta)}{\partial W} = v_0 \sigma(v^T W + c^T) - \sum_v p(v)\, v_0 \sigma(v^T W + c^T)}$$

where v is the vector $[v_1, v_2, \cdots, v_m]$ and W is an $m \times n$ matrix.

$\sigma(v^T W + c^T)$ is an element wise sigmoid of the vector inside

$$v_0 \sigma(v^T W + c^T) = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}_{m\times 1} \cdot \sigma\left([v_1\ v_2 \cdots v_m] \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix}_{m\times n} + [c_1\ c_2 \cdots c_n] \right)$$

Element wise Softmax operation

$$= \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix} \cdot \sigma\left(\left[\sum_{i=1}^{m} w_{i1} v_i + c_1,\ \sum_{i=1}^{m} w_{i2} v_i + c_2,\ \cdots \right.\right.$$
$$\left.\left. \cdots,\ \sum_{i=1}^{m} w_{in} v_i + c_n \right] \right)$$

$$= \begin{bmatrix} v_1 \cdot \sigma[z_1, z_2, \cdots, z_n] \\ v_2 \cdot \sigma[z_1, z_2, \cdots, z_n] \\ \vdots \\ v_m \sigma[z_1, z_2, \cdots, z_n] \end{bmatrix}_{m\times n} \Rightarrow \underset{p(H|V)}{E}[h_j v_i]$$

Now Let us derive gradients with respect to other variables :—

$$\frac{\partial L(\theta)}{\partial b_i} = \underset{p(H|V)}{E}[v_i] - \underset{p(H,V)}{E}[v_i] = \sum_H p(h|v)\, v_i - \sum_{V,H} p(h,v)\, v_i$$

$$= v_i \sum_H p(h|v) - \sum_V v_i\, p(v) \sum_H p(h|v)$$

$$\boxed{\frac{\partial L(\theta)}{\partial b_i} = v_i - \sum_V v_i\, p(v), \quad \frac{\partial L(\theta)}{\partial b} = v - \sum_v p(v)\, v}$$

Similarly $\frac{\partial L(\theta)}{\partial c_j} = \underset{p(H|V)}{E}[h_j] - \underset{p(H,V)}{E}[h_j]$

$$\frac{\partial L(\theta)}{\partial c_j} = p(h_j = 1|v) - \sum_v p(v)\, p(h_j = 1|v)$$

$$\boxed{\frac{\partial L(\theta)}{\partial c_j} = \sigma\left(\sum_{i=1}^{m} w_{ij} v_i + c_j \right) - \sum_v p(v) \sigma\left(\sum_{i=1}^{m} w_{ij} v_i + c_j \right)}$$

$$\nabla_c L(\theta) = \sigma(W^T v^{(0)} + c) - \mathop{E}_{p(v)}[\sigma(W^T v + c)]$$

Notice that all the gradient computations still involve this Expectation term, which is intractable. So we will set up a markov chain and sample to estimate these expectations.

RBM training with Block Gibbs Sampling :-

for all $v_d \in D$: $\rightarrow$ a sample $\rightarrow$ set of samples.

Randomly initialize $v^{(0)}$

for $t = 0, 1, \ldots k, k+1, \ldots, k+r$:

for $j = 1, 2, \ldots, n$

Sample $h_j^{(t)} \sim p(h_j | v^{(t)})$ $\Rightarrow$ given by $p(h_j = 1 | v)_m = \sigma\left(\sum_{i=1} w_{ij} v_i + c_j\right)$

for $i = 1, 2, \ldots, m$

Sample $v_i^{(t+1)} \sim p(v_i | h^{(t)})$

Notice that instead of going one by one from a sample to another, we make steps in blocks (first all hidden, then all visible) because we have either way to run the chain for large iterations. Hence each time step t actually consists of multiple time steps.

~~$W = W + \eta[\sigma(W$~~

$$W = W + \eta\left[v_d \sigma(v_d^T W + c^T) - \frac{1}{r}\sum_{p=k+1}^{k+r} v^{(p)} \sigma(v^{(p)T} W + c^T)\right]$$

notice that these terms are based on the actual sample v_d

these terms are based on the samples of the Markov Chain.

$$b = b + \eta\left[v_d - \frac{1}{r}\sum_{p=k+1}^{k+r} v^{(p)}\right]$$

$$c = c + \eta\left[\sigma(W^T v_d + c) - \frac{1}{r}\sum_{p=k+1}^{k+r} \sigma(W^T v^{(p)} + c)\right]$$

In practice, Gibbs Sampling can also be very inefficient because for every step of stochastic gradient descent, we need to run the Markov Chain for many steps.

In practice we use K-contrastive Divergence to train RBMs.
Contrastive Divergence uses the following idea:-
=> Instead of starting the Markov Chain at a random point, start from v_d where v_d is the current training instance.
=> Run Gibbs sampling for k steps, and denote the sample at the k^{th} step by $\tilde{v}$. Replace the expectation by a point estimate

$$E_{P(v,h)}[v_i h_j] \approx \sigma(W_j^T \cdot \tilde{v} + c_j) \cdot \tilde{v}_i$$

(W_j: j^{th} column of W)

Hence the gradients are as follows:-

$$W = W + \eta\left[v_d \cdot \sigma(v_d^T W + c^T) - \tilde{v} \cdot \sigma(\tilde{v}^T W + c^T)\right]$$

$$b = b + \eta\left[v_d - \tilde{v}\right]$$

$$c = c + \eta\left[\sigma(W^T v_d + c) - \sigma(W^T \tilde{v} + c)\right].$$

The higher the value of k, the less biased the estimate of the gradient will be.

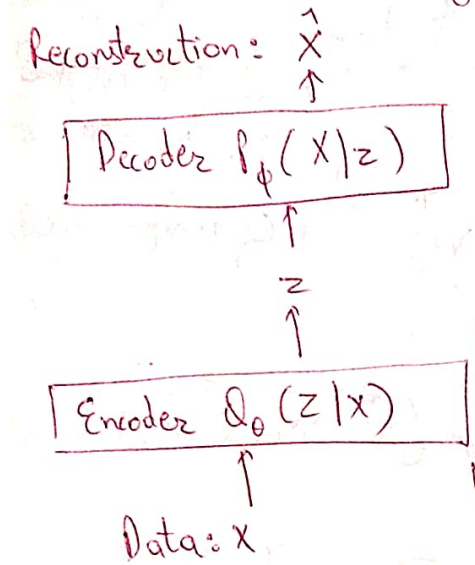
Variational Autoencoders:-

Before VAE's some info about autoencoders. An autoencoder contains an encoder which takes the input X and maps it to a hidden representation, then the decoder takes this hidden representation and tries to reconstruct the input from it as $\hat{X}$. The loss between $\hat{X}$ and X is used to train the params of the network (Unsupervised task). Can we use Autoencoders for generative purposes? well given a hidden representation, we can produce an $\hat{X}$. But it is a deterministic procedure so variability will be less. Moreover, out of all the possible hidden representations, only a few are likely to be produced by our data. So deciding a good representation (hidden) for generation is still a question. In case of RBM's we have and approximate for the joint dist so we can find the likely hidden representations.

Variational auto encoders have similar/same structure as autoencoders, but they learn a distribution over the hidden variables.

visible variables will be referred as X, and hidden variables will be referred as z.

The architechture and goals are as follows :-



Goal 1 :- Learn a Distribution over the latent variables $Q(z|X)$.

Goal 2 :- Learn a distribution over the visible variables $P(X|z)$

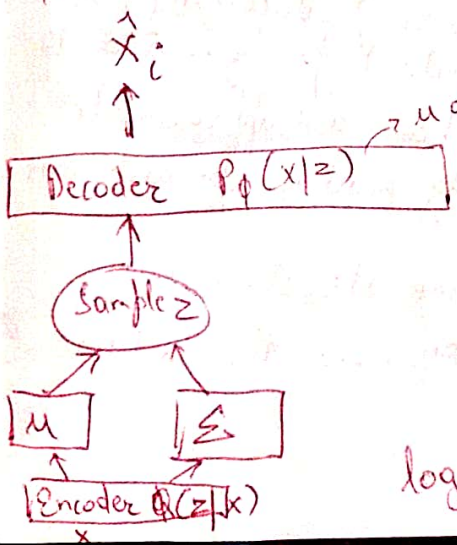
Notice that both the encoder and decoder learn distributions rather than the actual hidden / visible representations.

θ : the parameters of the encoder network.

ϕ : the parameters of the decoder network.

In the encoder part, we assume that the latent variables come from a standard normal distribution $N(0, I)$ and the job of the encoder is to predict the parameters of this distribution. The input of the encoder is the visible variables X, and it has to give the parameters (μ, Σ) of the distribution $Q_\theta(z|X) \sim N(\mu, \Sigma)$, and the reference distribution is $N(0, I)$

The job of the decoder is to predict a probability over X : $P(X|z)$ We assume that $P(X|z)$ is a Gaussian Distribution with unit variance. The decoder predicts the mean of the distribution, $P_\phi(X|z)$.



The loss function is as follows :- we want to maximize the likelihood the probability of the data. Basically

$$P(x_i) = \int P(z)\, P(x_i|z)\, dz \quad \text{multivariate integral.}$$

$$= E_{z \sim Q_\theta(z|x_i)} [\log(P_\phi(x_i|z))]$$

$P(z)$ only.

log is taken for numerical stability.

But we want to also ensure that $Q_\theta(z|x_i)$ does not diverge much from $N(0, I)$. So a KL divergence term between these two distributions is also taken. So the total loss :—

$$L(x_i) = -\mathbb{E}_{z \sim Q_\theta(z|x_i)}[\log(P(x_i|z))] + KL(Q_\theta(z|x_i) \,||\, N(0,I))$$

We need to add this divergence term so that the network does not cheat. Otherwise it might just learn a unique representation for each datapoint and decode it as a mapping without learning anything. So we predict a distribution $Q_\theta(z|x_i)$ such that a sample from this distribution should be able to reconstruct the original data point with high probability. We assume that $P(z) \sim N(0, I)$ because any distribution in d dimension may be generated with a set of d normally distributed variables by using a sufficiently complex function. So the decoder learns this function to arrive at $P_\phi(x_i|z)$ using only normally distributed variables.

It can be shown that $KL[N(\mu(x), \Sigma(x)) \,||\, N(0,1)]$

$$= \frac{1}{2}\left(tr(\Sigma(x)) + \mu(x)^T[\mu(x) - k - \log\det(\Sigma(x))]\right)$$

where $\Sigma(x)$ and $\mu(x)$ are the μ and Σ calculated by the decoder for $Q_\theta(z|x_i)$, it is expressed like this since it is a function of X. k is the dimensionality of the latent variables. This is easily computable. But $\mathbb{E}_{z \sim Q_\theta(z|x)}[\log P(x|z)]$ is still a bottleneck since it involves a multivariate integral. But in VAE's we approximate this expectation using a single z sampled from $N(\mu(x), \Sigma(x))$. Hence $\log P(x|z) = C - \frac{1}{2}\|x_i - \mu(z)\|^2$

$\mu(z)$ is the mean of the distribution $P_\phi(x|z)$ calculated by the decoder. It is of this form as $\mu(z)$ depends on z, moreover the expression can be derived from the PDF of Normal distribution

↳ we assume $P_\phi(x|z)$ is normal with unit variance.

Hence the effective objective function is :—

$$\underset{\theta,\phi}{\text{minimize}} \sum_{n=1}^{N}\left[\frac{1}{2}\left[tr(\Sigma(x)) + \mu(x)^T[\mu(x) - k - \log\det(\Sigma(x))\right] + \|x - \mu(z)\|^2\right]$$

Observe that $\mu(x)$ and $\Sigma(x)$ depend on θ, while $\mu(z)$ depends on ϕ.

But since the network involves sampling z. It is not end to end differentiable. To make it differentiable we sample ϵ in the input step itself. So we calculate z like this :-

Sample $\epsilon \sim N(0, I)$ in the input stage. Then

$$z = \mu(x) + A(x)\epsilon$$, where $A(x)$ is such that $\Sigma(x) = A(x)A(x)^T$

This is because z will now follow $N(\mu(x), AIA^T)$

$$= N(\mu(x), \Sigma(x))$$

Now the parameters θ and ϕ can be learnt using gradient descent and backpropogation.

In abstraction, we calculate z in the above way.

In Generation, we sample $z \sim N(0, I)$ and just feed it to the decoder, since model was trained to converge $Q_\theta(z|x) \sim N(0, I)$

Autoregressive Models :- these models do not contain any latent variables They also learn a joint distribution over X. For us, $X \in \{0,1\}^n$

AR models do not make any independence assumptions but use the default factorization of $p(x)$:- $p(x) = \prod_{i=1}^{n} p(x_i | x_{<i})$

$$= p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_2, x_1) \cdots p(x_n|x_{<n})$$

In AR models, these factors are parameterized using a neural network and these parameters are learnt to eventually learn the factors which are used to construct the joint distribution.

The network is set up such that, at the output layer, we predict the n factors/conditional probability distributions. The i^{th} neuron in the output layer gives $p(x_i = 1 | x_{<i})$. But the network should be constructed such that i^{th} output should only be connected to the previous $i-1$ inputs. Since it only depends on these $i-1$ variables.

In NADE (Neural Autoregressive Density Estimator), a hidden representation is calculated using only the $i-1$ inputs, then this hidden representation is used to calculate $p(x_i | x_{<i})$. So at the i^{th} step :-

$$\boxed{h_i = \sigma(W_{<i} x_{<i} + b)}$$ where $h_i \in \mathbb{R}^d$ and $W \in \mathbb{R}^{d \times n}$, $b \in \mathbb{R}^d$

So only first $i-1$ columns are used to calculate h_i. Then

$y_i = p(x_i | x_{<i})$ where $\boxed{y_i = (V_i h_i + c_i)}$ $\quad V_i \in \mathbb{R}^d$, $c_i \in \mathbb{R}$

Notice that for calculating h_1, no input is seen, hence we make h_1 as a learnable parameter of the network. So in total there are :- $d \times n$ (for W) + d (for b) + $n \times d$ (for V_is) + n (for c_is)

this quantity is linear wrt n, instead of exponential. This is because we use weight tying. We use the same parameters W and b at each step. In practice, we mask the input appropriately to select the correct columns of W.

=) We can use the cross entropy loss for each node at the output as the loss for training. Since each output node predicts a distribution of a binary variable we can calculate its cross entropy loss and take the sum of all the losses coming from each node.

=) Notice that these models are not meant for abstraction by design. This is because we don't get a hidden representation for the whole input together from the models. We can of course devise a way to merge the individual hidden representation to calculate an abstraction

=) For generation, first the model gives $p(x_1)$, using this Bernoulli dist, we can sample x_1, and use the sampled value to calculate h_2 and then $y_2 (p(x_2|x_1))$. Then x_2 can be sampled and then x_1 and x_2 are used to calculate y_3. This process is repeated for all n vars.

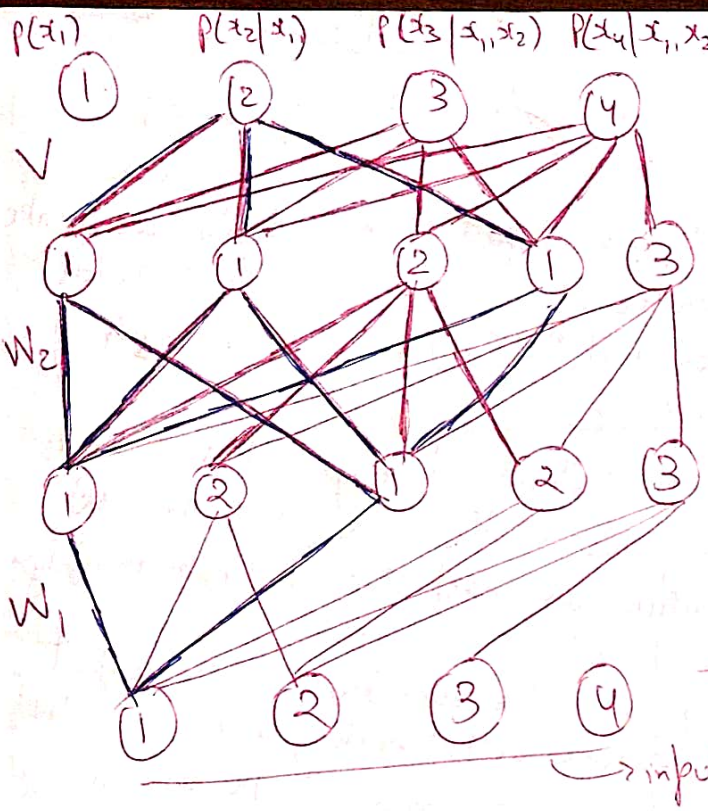
This process is sequential and slow. However it can be sped up by reusing calculations made in the previous step :-

$$W_{<i} x_{<i} = \underbrace{W_{<i-1} x_{<i-1}}_{\text{already calculated in previous step}} + W_{:i-1} x_{i-1}$$

$W_{:i-1}$ is the $i-1^{th}$ col of W and x_{i-1} is a scalar.

Another autoregressive model is MADE (Masked Autoencoder Density Estimator) which also considers the natural factorization. It uses the same architecture as an autoencoder, but introduces masks in the hidden layer computations which makes sure that the final i^{th} output depends on the previous $i-1$ inputs.

We start by assuming some ordering on the inputs and just number them from 1 to n.



the output only accepts input from nodes numbered strictly less that it self.

the hidden layer accepts input from nodes numbered less than or equal to itself.

Observe the blue weights. For output 2, the initial input is only ①, which is made sure by this network.

hidden.

⇒ Each node is randomly assigned a number between 1 to n-1

⟹ only for understanding, in practices, only masking is done.

Let ~~masked~~ $W_2 =$

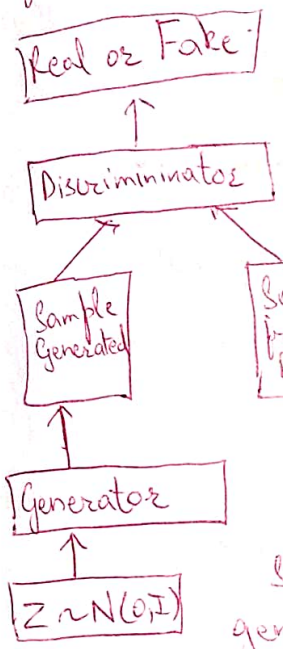
$$\begin{bmatrix} w_{11}^{(2)} & w_{12}^{(2)} & \cdots & w_{15}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & \cdots & w_{25}^{(2)} \\ \vdots & \vdots & & \vdots \\ w_{51}^{(2)} & w_{52}^{(2)} & \cdots & w_{55}^{(2)} \end{bmatrix}_{5\times 5} \odot \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \cancel{z_1} \\ z_2 \\ \vdots \\ z_5 \end{bmatrix}$$

W_2. mask. ↑ input from first hidden layer

Since the first node of 2nd hidden layer depends only on 1st and 3rd node of 1st hidden layer, hence the mask contains [1 0 1 0 0] as the first row. The 3rd row of 2nd hidden layer depends only on first 4 nodes of 1st hidden layer. So the 3rd row of the mask is [1 1 1 1 0].

The loss function is same sum of cross entropy loss, and it follows the same autoregressive generation process as in ~~NADE~~

Generative Adversarial Networks:- Instead of learning a joint distribution over the data random variables, GANs take a different approach to this problem where the idea is to sample from a simpler tractable distribution; say $z \sim N(0, I)$; and then learn a complex transformation from this to the training distribution. The training of the network is like a two player game.



The job of the generator is to produce images which look natural as if sampled from the real dataset and fool the discriminator.

The job of the discriminator is to get better and better at distinguishing between true images and generated (fake) images.

Let G_ϕ be the generator and D_θ be the discriminator. G_ϕ takes a random noise vector $z \sim N(0,I)$ and produces x, $G_\phi(z) = x$. Let the discriminator output a score between 0 and 1. For a given z, the generator would want to maximize $\log D_\theta(G_\phi(z))$ or minimize $\log(1 - D_\theta(G_\phi(z)))$. Since we want to minimize this loss over all z, we minimize the expectation of this loss. So for the generator

$$\min_\phi \; \mathbb{E}_{z \sim N(0,I)}\left[\log(1 - D_\theta(G_\phi(z)))\right]$$

The discriminator on the other hand has to assign a high score to real images and a low score to generated images. It should do for all possible real images and all possible fake images.

$$\max_\theta \; \mathbb{E}_{x \sim p_{data}}[\log D_\theta(x)] + \mathbb{E}_{z \sim N(0,I)}[\log(1 - D_\theta(G_\phi(z)))],$$ combining both the objectives of the generator and discriminator:-

$$\min_\phi \max_\theta \; \mathbb{E}_{x \sim p_{data}}[\log D_\theta(x)] + \mathbb{E}_{z \sim N(0,I)}[\log(1 - D_\theta(G_\phi(z)))]$$

so the overall training proceeds by alternating between these two steps:-

i) Gradient Ascent on Discriminator
ii) Gradient Ascent on Generator.

In practice, the above generator objective does not work well and we use a different objective:-
we minimize $-\log(D_\theta(G_\phi(z)))$.

GAN Training :-
for number of iterations.
for k steps
Sample m noise samples $\{z^{(1)}, z^{(2)}, \cdots, z^{(m)}\}$
Sample m examples $\{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$ from data.
update the discriminator by ascending its gradient:-

$$\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} [\log D_\theta(x^{(i)}) + \log(1 - D_\theta(G_\phi(z^{(i)})))]$$

Sample m noise samples $\{z^{(1)}, z^{(2)}, \cdots, z^{(m)}\}$
X ~~Sample m examples $\{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$ from data~~.
update the generator by ascending its gradient:-

$$\nabla_\phi \frac{1}{m} \sum_{i=1}^{m} \log D_\theta(G_\phi(z^{(i)}))$$

Architecture :- For discriminator any CNN based classifier with 1 class at the output can be used (eg: VGG, ResNet).

For the Generator, we have to go from a random vector to an image, called transpose convolution.
Pooling layers are replaced by strided convolutions (discriminator) and fractional-strided convolutions (generator). ReLU is used in generator except at output where tanh is used as activation function. Use LeakyReLU in the discriminator for all layers.

It can be formally proved that when solve the above objective, the distribution learned by the Generator over the variables is same as the ~~distribution~~ actual distribution.

Theorem:- The global minimum of the virtual training criterion $C(G) = \max_D V(G,D)$ is achieved if and only if $p_g = p_{data}$.

So even though GANs do not model $P(X)$, they implicitly learn $P(X)$ through an indirect modified objective.

Summary:-

	RBMs	VAEs	AR models	GANs
Abstraction	Yes	Yes	No	No
Generation	Yes	Yes	Yes	Yes
Compute $P(x)$	Intractable	Intractable	Tractable	No
Sampling	Gibbs Sampling	Fast	Slow	Fast
Type of GM	Undirected	Directed	Directed	Directed
Loss	KL Divergence	KL Divergence	KL Divergence	Jensen-Shannon Divergence
Assumptions	X independent given z	X independent given z	None	Not applicate since $P(x)$ is not computed.
Samples	Bad	OK	Good	Good (best)

Recent works combine these models eg Adversarial Autoencoders, Pixel GAN Autoencoders, etc.