ECE499/ECE590 Machine Learning for Embedded Systems (Fall 2021)

Lecture 2: Train Neural Networks

Weiwen Jiang, Ph.D.

Electrical and Computer Engineering

George Mason University

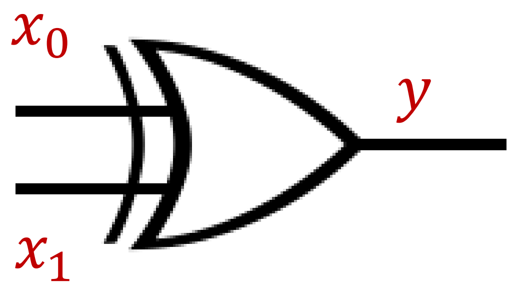
wjiang8@gmu.edu

Artificial Neuron Design

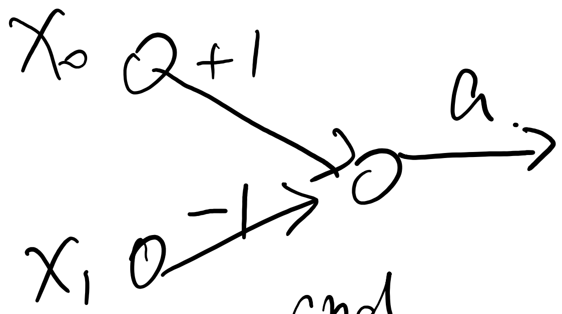
- **Idealized neuron models**
  - Idealization removes complicated details that are not essential for understanding the main principles.
  - It allows us to apply mathematics and to make analogies.
- **Break the limitations on MP Neuron**
  - What about non-boolean inputs (say, real number)? ✓
  - What if we want to assign more weight (importance) to some inputs? ✓
  - What about functions which are not linearly separable ? ? **=> MLP**
  - Do we always need to hand code the threshold? ? **=> Training**

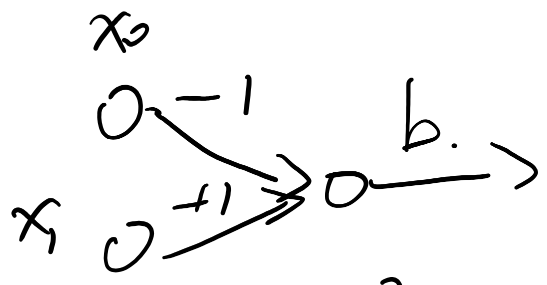
Multi-Layer Perceptron (MLP)

Solving



$$= \underbrace{(x_0 \; and \; \overline{x_1})}_{a.} \; or \; \underbrace{(\overline{x_0} \; and \; x_1)}_{b.}$$



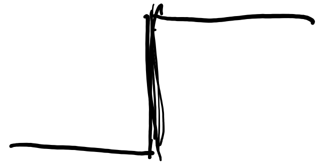


$\theta = 2$.

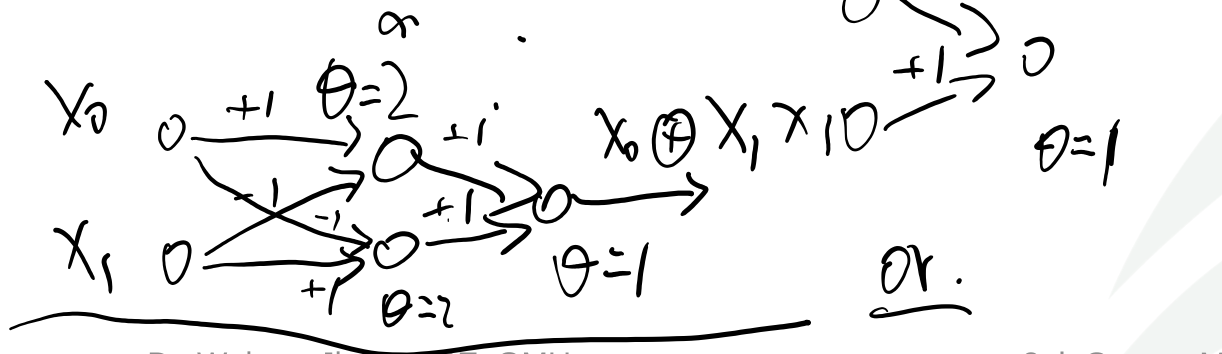
$x_0 x_1$	y	z_1
0 0	0	0
0 1	1	1
1 0	1	1
1 1	1	2

$\theta = 1$

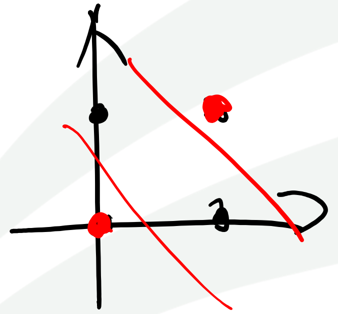
and



$1 < \theta < 2 \quad \underline{\theta = 2.}$



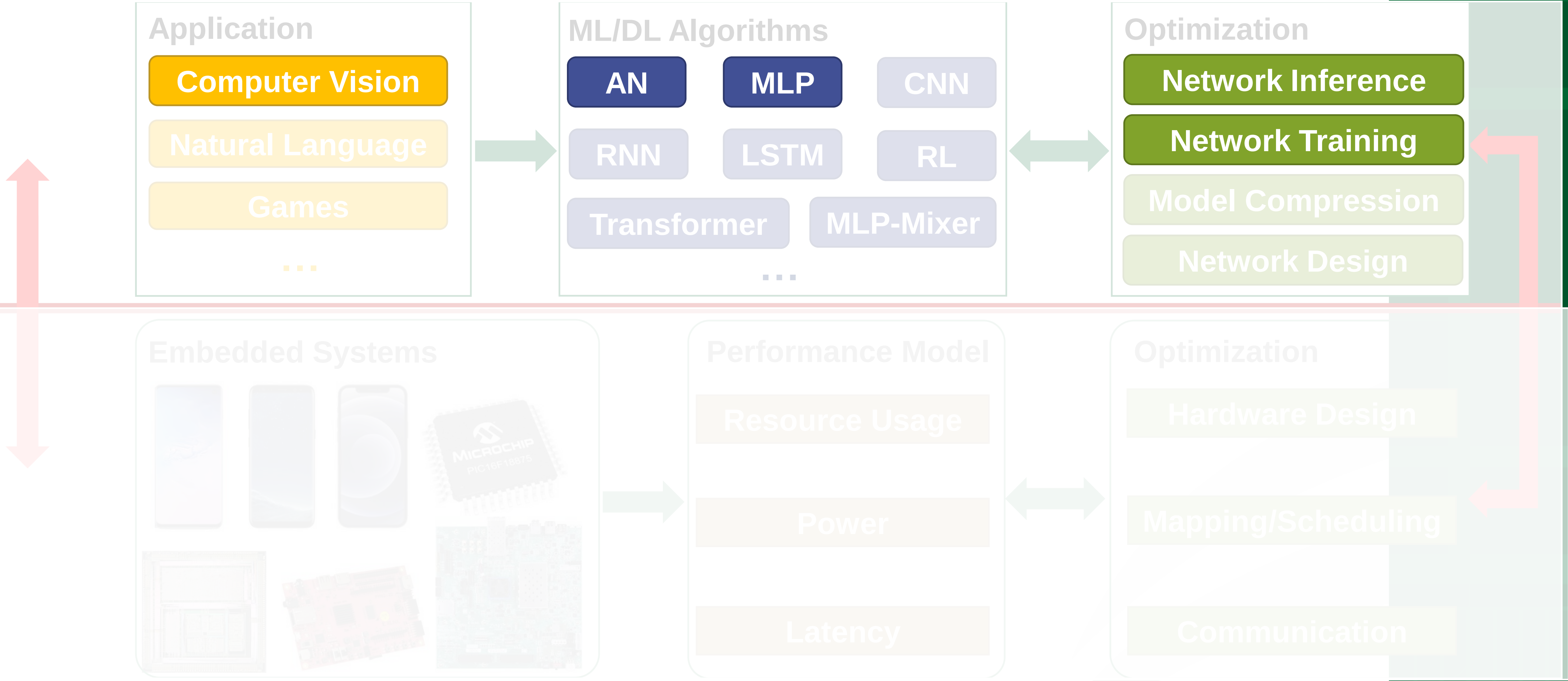
$\underline{or.}$



Artificial Neuron Design

- **Idealized neuron models**
  - Idealization removes complicated details that are not essential for understanding the main principles.
  - It allows us to apply mathematics and to make analogies.
- **Break the limitations on MP Neuron**
  - What about non-boolean inputs (say, real number)? ✓
  - What if we want to assign more weight (importance) to some inputs? ✓
  - What about functions which are not linearly separable ? ✓
  - Do we always need to hand code the threshold? ? **=> Training**

Week 2: From Inference to Training

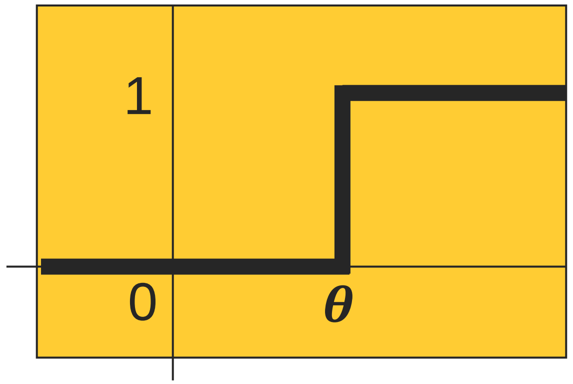


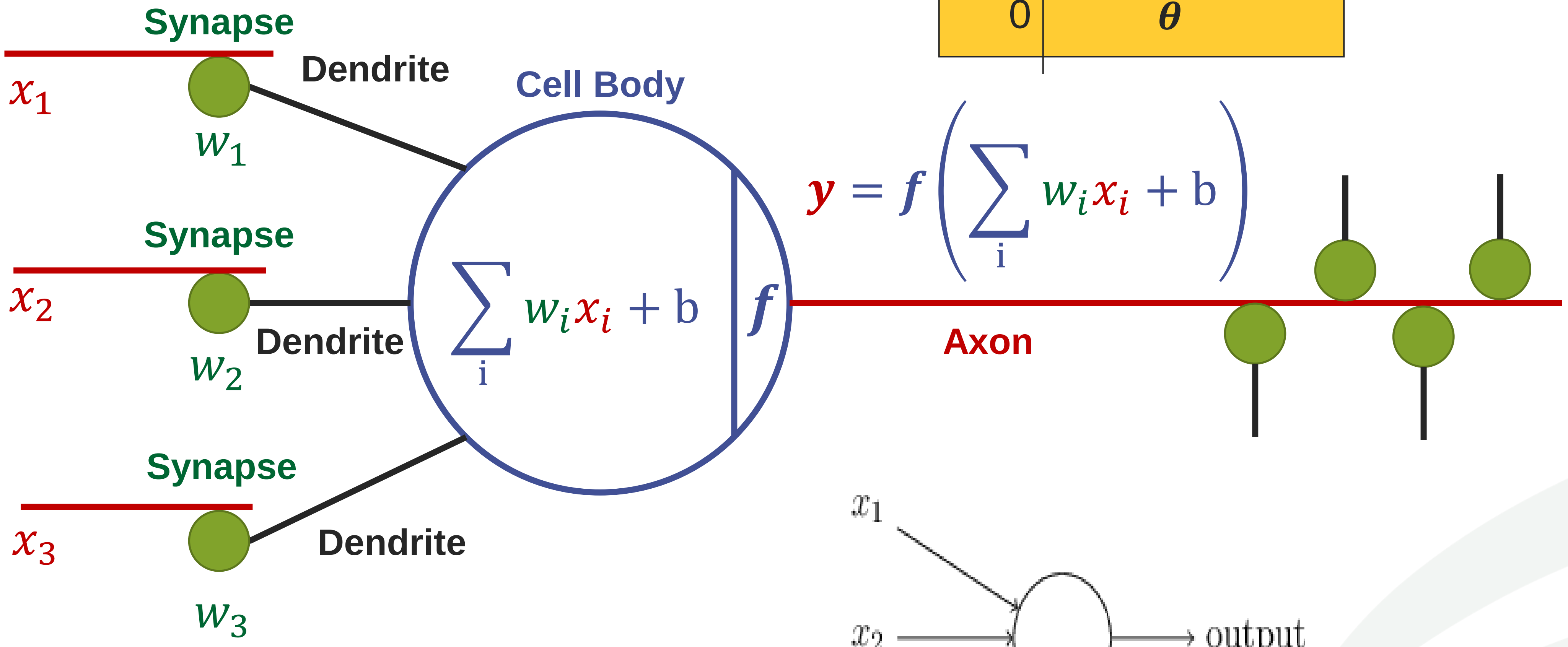
Agenda

- Artificial neuron network for multi-class classification
- Inference: forward propagation
- Training: backpropagation
- Finite-differencing and backpropagation
- Cross-entropy function as objective
- Overfitting issues of neural network
- Regularization techniques to overcome overfitting
- Initialization and softmax output layer
- Conclusions

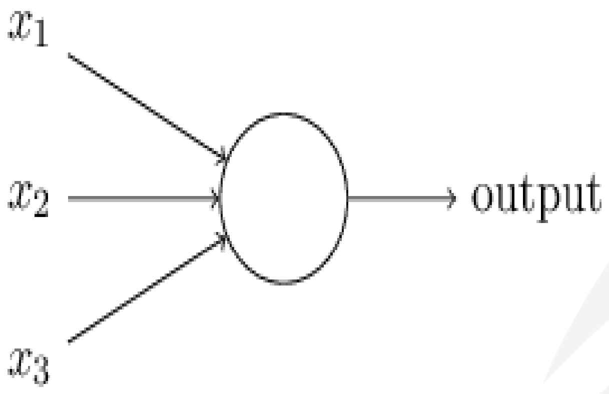
Perceptron

Frank Rosenblatt @ 1958





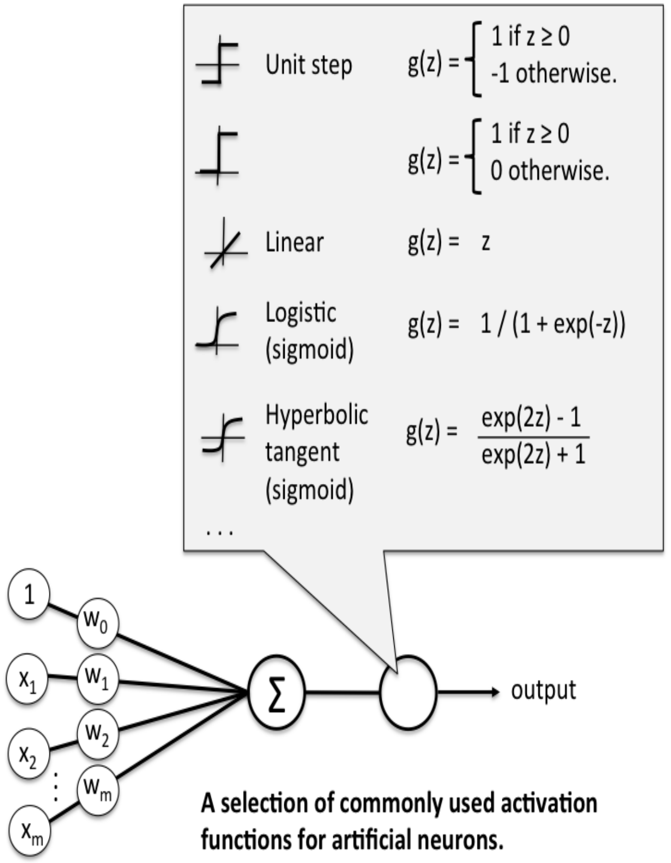
$$y = f\left(\sum_i w_i x_i + b\right)$$



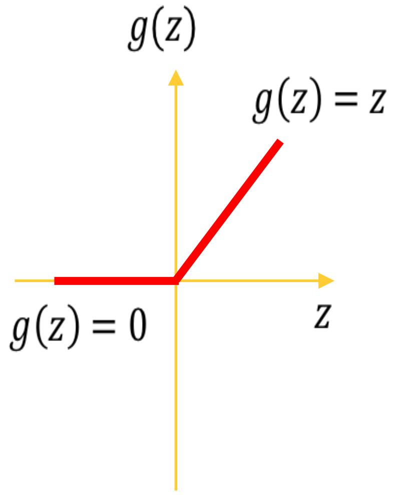
Various activation functions

- People invent new activation functions and publish papers

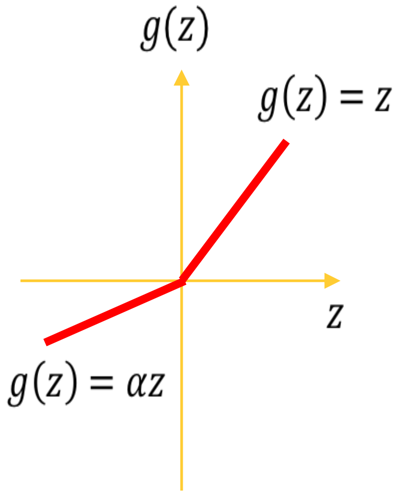
"Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," K. He et al., ICCV 2015


Unit step
g(z) = 1 if z ≥ 0; -1 otherwise.
g(z) = 1 if z ≥ 0; 0 otherwise.
Linear
g(z) = z
Logistic (sigmoid)
g(z) = 1 / (1 + exp(-z))
Hyperbolic tangent (sigmoid)
g(z) = (exp(2z) - 1) / (exp(2z) + 1)
...
1
x1
x2
xm
w0
w1
w2
wm
Σ
output

A selection of commonly used activation functions for artificial neurons.



Rectified Linear (ReLU) activation function

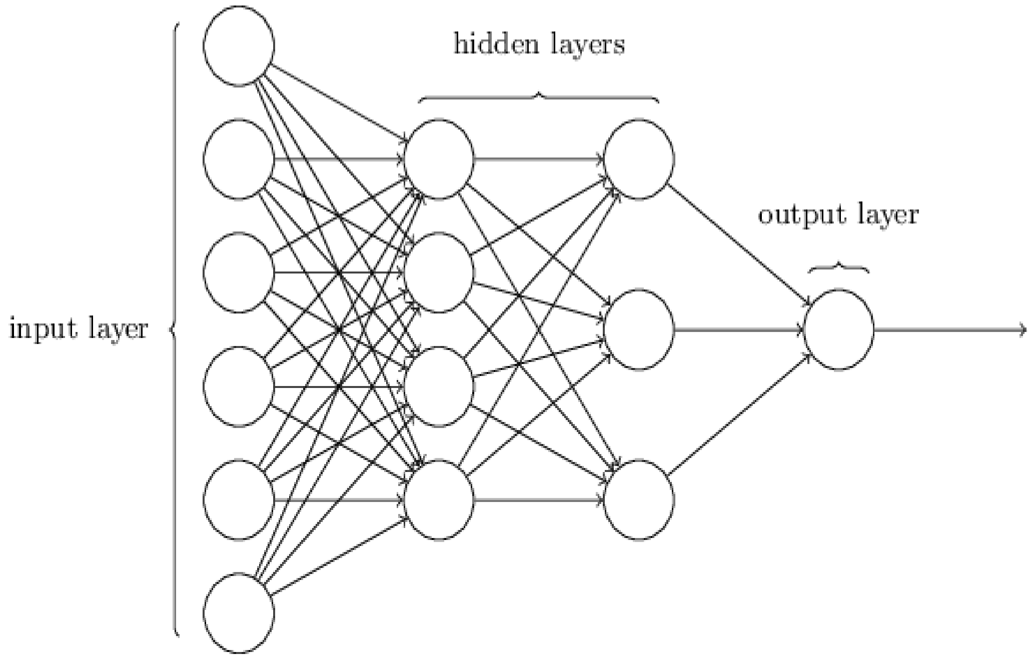


Parameterized Rectified Linear (PReLU) activation function

Latest research progress on defining new activation functions

Neural network terminologies

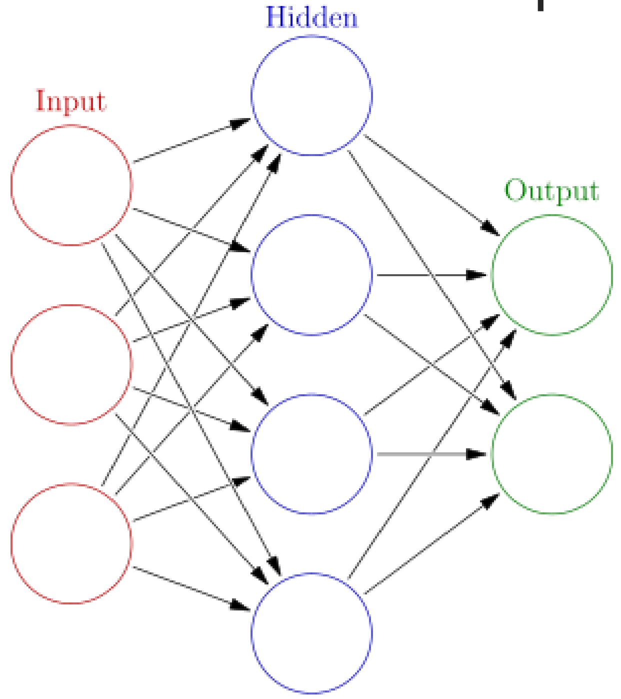
- Input layer, output layer and hidden layers



For historical reasons, such multiple layer networks are sometimes called multilayer perceptrons or MLPs, though they are mostly made up of sigmoid neurons, not perceptrons

Neural network for multi-class classification

- Since our neuron (such as sigmoid neuron) is designed for (binary) classification problem, the so-built neuron network would be naturally suitable for solving (binary) classification problems
- Moreover, there is nothing to stop us from having multiple outputs, so it can be used for multi-class classification problems

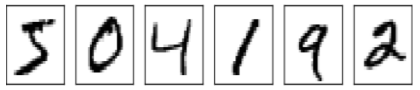


An example for multi-class classification formulation

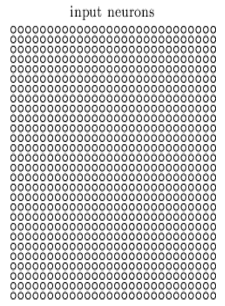
- Classify images of handwriting digits to 10 classes

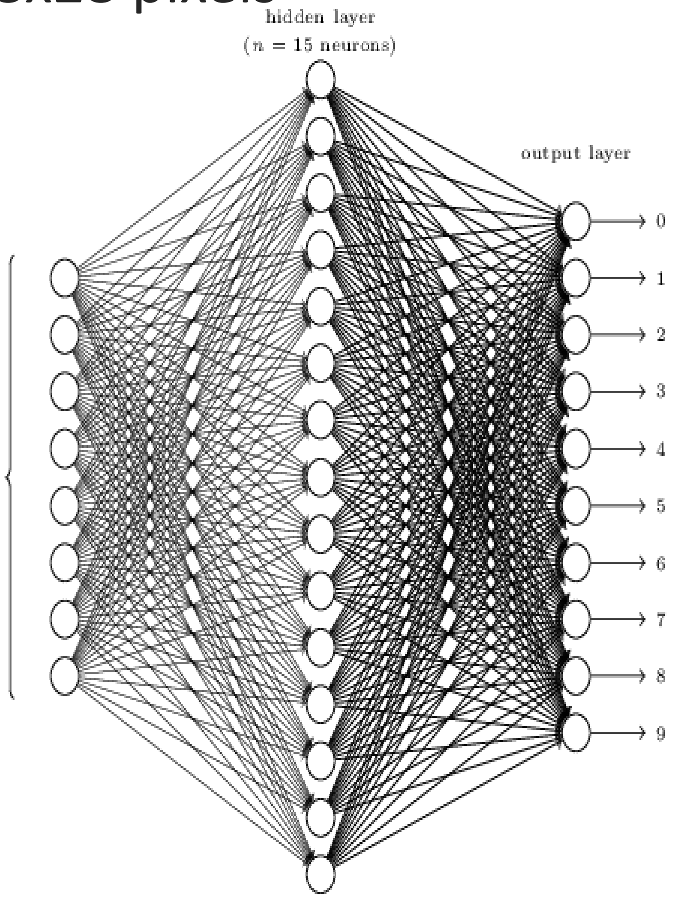
  Image size = 28x28 pixels

$x^{(i)}$



28x28 pixels





$y^{(i)}$

$$\begin{bmatrix}0\\0\\0\\0\\0\\1\\0\\0\\0\\0\end{bmatrix}\quad\begin{bmatrix}1\\0\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}\quad\begin{bmatrix}0\\0\\0\\0\\1\\0\\0\\0\\0\\0\end{bmatrix}\quad\begin{bmatrix}0\\1\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}\quad\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\0\\1\end{bmatrix}\quad\begin{bmatrix}0\\0\\1\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}$$

Agenda

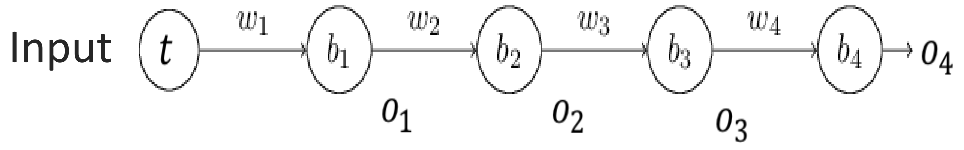
- Artificial neuron network for multi-class classification
- Inference: forward propagation
- Training: backpropagation
- Finite-differencing and backpropagation
- Cross-entropy function as objective
- Overfitting issues of neural network
- Regularization techniques to overcome overfitting
- Initialization and softmax output layer
- Conclusions

Forward propagation

- Given a neural network model with weights and biases (parameters), forward propagation computes the outputs, starting from the inputs

  A simple example:

  Input to each neuron: $z = b + wx$ Neuron activation function: $g(z)$



$$o_1 = g(z_1) = g(b_1 + w_1 t)$$

$$o_2 = g(z_2) = g(b_2 + w_2 o_1)$$

$$o_3 = g(z_3) = g(b_3 + w_3 o_2)$$

$$o_4 = g(z_4) = g(b_4 + w_4 o_3)$$

  This is also called "inference"

- The same procedure holds for the more complicated neural networks too

  The output class is determined by the maximum value among all output neurons

Agenda

- Artificial neuron network for multi-class classification
- Inference: forward propagation
- Training: backpropagation
  - Math
  - A quick review of unconstrained optimization
  - SGD and backpropagation
- Finite-differencing and backpropagation
- Cross-entropy function as objective
- Overfitting issues of neural network
- Regularization techniques to overcome overfitting
- Initialization and softmax output layer
- Conclusions

Derivative of a function

- The first order derivative of a function at point $(x, f(x))$ is defined as the instantaneous rate of change at that point, which is given by the limit of the average rate of change as

$$\frac{df(x)}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

- Some examples
  - Power rule $\frac{d}{dx}(x^p) = px^{p-1}$
  - Exponential rule $\frac{d}{dx}(b^x) = b^x \ln(b)$ $\frac{d}{dx}(e^x) = e^x$
  - Logarithm rule $\frac{d}{dx}(log_b(x)) = \frac{1}{x}\frac{1}{\ln(b)}$ $\frac{d}{dx}(ln(x)) = \frac{1}{x}$
  - Constant rule $\frac{d}{dx}(C) = 0$

Properties of derivatives

- Constant

$$\frac{d}{dx}(cf(x)) = c\frac{d}{dx}(f(x))$$

- Sum and subtraction

$$\frac{d}{dx}(f(x) \pm g(x)) = \frac{d}{dx}(f(x)) \pm \frac{d}{dx}(g(x))$$

- Product

$$\frac{d}{dx}\big(f(x)g(x)\big) = \frac{d}{dx}\big(f(x)\big)g(x) + f(x)\,\frac{d}{dx}(g(x))$$

- Division

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{\frac{d}{dx}\big(f(x)\big)g(x) - f(x)\,\frac{d}{dx}(g(x))}{g(x)^2}$$

- Chain rule

$$\frac{d}{dx}(f(g(x))) = \frac{d}{dx}\big(f(g(x))\big)\frac{d}{dx}(g(x))$$

So far the function and derivatives are defined for scalars

- Scalars: 1 dimensional data
- Now let’s extend the same concept to

  Function of a vector of variables

  A vector function of a vector of variables

Gradient

- Let f be a real-valued function of n variables

$$f(x) = f(x_1, x_2, \dots,, x_n)$$

- Gradient is defined as a vector of first derivatives
- Hessian is defined as a matrix of second derivatives (which is also symmetric)

$$\nabla f(x) \equiv \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \frac{\partial^2 f}{\partial x_2\,\partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n\,\partial x_1} & \frac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

An example of Gradient & Hessian

$$f(x) = f(x_1, x_2, x_3) = \frac{1}{2}(4x_1^2 + 4x_1x_2 + 2x_1x_3 + 5x_2^2 + 6x_2x_3 + 7x_3^2) + 2x_1 - 8x_2 + 9x_3$$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \frac{\partial f(x)}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 4x_1 + 2x_2 + x_3 + 2 \\ 2x_1 + 5x_2 + 3x_3 - 8 \\ x_1 + 3x_2 + 7x_3 + 9 \end{bmatrix}$$

$$H = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 x_2} & \frac{\partial^2 f(x)}{\partial x_1 x_3} \\ \frac{\partial^2 f(x)}{\partial x_2 x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \frac{\partial^2 f(x)}{\partial x_2 x_3} \\ \frac{\partial^2 f(x)}{\partial x_3 x_1} & \frac{\partial^2 f(x)}{\partial x_3 x_2} & \frac{\partial^2 f(x)}{\partial x_3^2} \end{bmatrix} = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 7 \end{bmatrix}$$

An example of Gradient & Hessian (continued)

- You may have noticed the previous example is a quadratic form

$$f(x) = f(x_1, x_2, x_3) = \frac{1}{2}(4x_1^2 + 4x_1x_2 + 2x_1x_3 + 5x_2^2 + 6x_2x_3 + 7x_3^2) + 2x_1 - 8x_2 + 9x_3$$
$$= \frac{1}{2}(4x_1^2 + 2x_1x_2 + x_1x_3 + 2x_1x_2 + 5x_2^2 + 3x_2x_3 + x_1x_3 + 3x_2x_3 + 7x_3^2) + 2x_1 - 8x_2 + 9x_3$$
$$= \frac{1}{2}x^TQx + b^Tx$$

$$Q = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 7 \end{bmatrix} \qquad b = \begin{bmatrix} 2 \\ -8 \\ 9 \end{bmatrix}$$

$$\nabla f(x) = Qx + b = \begin{bmatrix} 4x_1 + 2x_2 + x_3 + 2 \\ 2x_1 + 5x_2 + 3x_3 - 8 \\ x_1 + 3x_2 + 7x_3 + 9 \end{bmatrix}$$

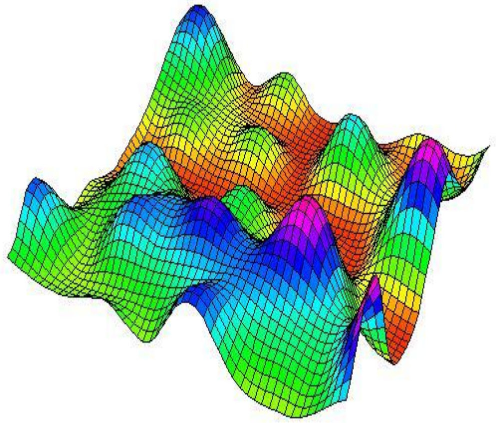
$$H = Q = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 7 \end{bmatrix}$$

Agenda

- Artificial neuron network for multi-class classification
- Inference: forward propagation
- Training: backpropagation
  - Math
  - A quick review of unconstrained optimization
  - SGD and backpropagation
- Finite-differencing and backpropagation
- Cross-entropy function as objective
- Overfitting issues of neural network
- Regularization techniques to overcome overfitting
- Initialization and softmax output layer
- Conclusions

Unconstrained optimization problem

$$\text{Minimize}_{x \in S}\ f(x)$$



- No fundamental difference between minimization and maximization problems

$$\text{Maximize}_{x \in S}\ f(x) \equiv \text{Minimize}_{x \in S}\ -f(x)$$

- For optimization, a first-order necessary condition for a local minimizer is given by

$$\nabla f(x) = 0$$

General optimization algorithm

- Specify some initial guess of the solution x(0)
- For k=0,1,...
  - If x(k) is optimal, stop
  - Determine an improved estimate of the solution

$$x(k+1) = x(k) + \alpha(k) * p(k)$$

$\alpha(k)$: *a scalar for step length*
$p(k)$: *a search direction*

- Most practical optimization algorithms follow this paradigm
- It is an iterative algorithm
- The computed "solution" is only an approximation
  - Stop condition: typically when the value of f(k), x(k) and/or p(k) changes little or some approximation of the first-order condition ($\nabla f(x) = 0$) is satisfied

Steepest descent method vs Gradient descent method

- The simplest method with a descent direction defined as

$$p(k) = -\nabla f\big(x(k)\big)$$

–This is indeed a descent direction

Taylor series approximation for $f(x)$: $\quad f(x(k)+p(k)) \approx f\big(x(k)\big) + \nabla f\big(x(k)\big)p(k)$

$$\nabla f\big(x(k)\big)p(k) = -\nabla f\big(x(k)\big)\nabla f\big(x(k)\big) < 0$$

- Step length to update search

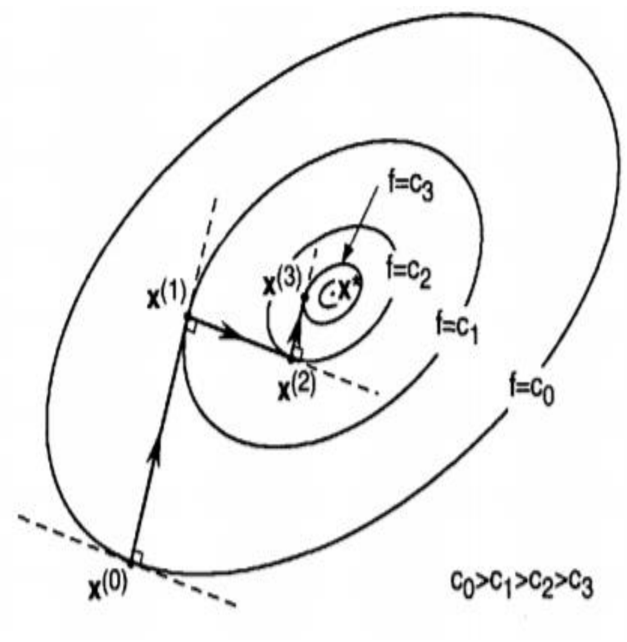
$$x(k+1) = x(k) + \alpha(k) * p(k)$$

–Steepest descent

$$a(k) = argmin_{a(k)>0}\Big(f\big(x(k) + \alpha(k) * p(k)\big)\Big)$$

–Gradient descent

$$a(k) > 0 \text{ of some sort}$$



Objective function in a special form of a sum

- Most machine learning problems consider an objective in a form of a sum

$$f(w) = \frac{1}{N}\sum_{i=1}^{N} f_i(w)$$

where $f_i(w)$ is typically related to the i–th observation in the data set

–E.g., regression objective

$$\text{Minimize}_w \; f(w) = \frac{1}{2}\sum_{i=1}^{n}\left(y^{(i)} - H\left(w, x^{(i)}\right)\right)^2$$

where n is the number of total observations used for training

- The gradient computation depends on the size of data set N

$$\nabla f(w) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(w)$$

–When N is large (for Big Data application), the computation cost becomes too high

Agenda

- Artificial neuron network for multi-class classification
- Inference: forward propagation
- Training: backpropagation
  - Math
  - A quick review of unconstrained optimization
  - SGD and backpropagation
- Finite-differencing and backpropagation
- Cross-entropy function as objective
- Overfitting issues of neural network
- Regularization techniques to overcome overfitting
- Initialization and softmax output layer
- Conclusions

Stochastic gradient descent (SGD)

- Instead of using all data set N to compute the standard (or “batch”) gradient descent, SGD uses samples of a subset of N summand functions at every iteration

$$\nabla f(w) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(w) \approx \frac{1}{N_m}\sum_{i=1}^{N_m} \nabla f_i(w) \qquad \text{with } N_m \ll N$$

“Stochastic” means the algorithm “randomly” samples the data at every iteration

This is also called mini-batch SGD

- On-line (confusedly, also called stochastic) gradient descent uses one sample at a time

$$\nabla f(w) \approx \nabla f_i(w)$$

Training a neural network

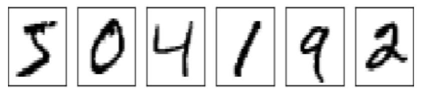
- The model parameters are the weights (and biases) of the model

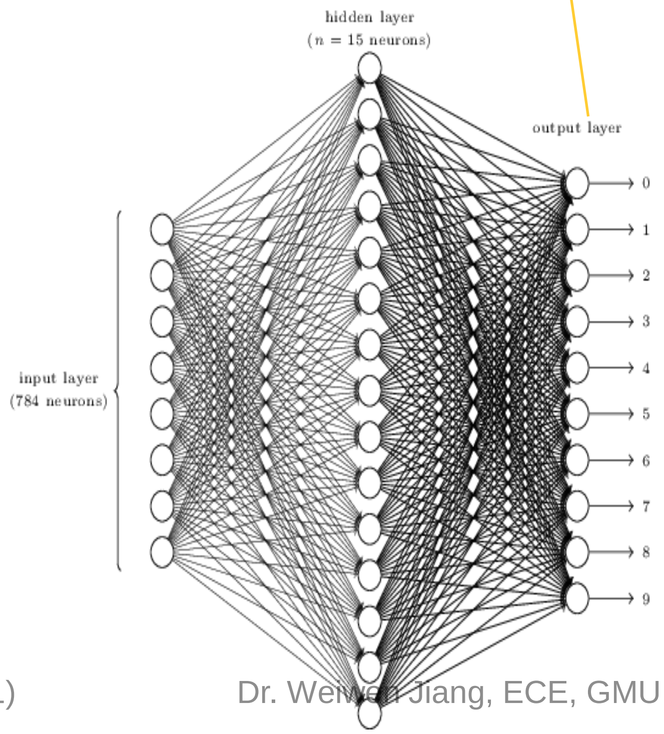
Input to each neuron: $z = w_0 + w_1 x_1 + \cdots + w_m x_m = \sum_{j=0}^{m} w_i x_i = w^T x$

- One objective can be to minimize the SSE

$$\min_w : f(w) = \frac{1}{2N} \sum_{i=1}^{N} \left(y^{(i)} - h\left(w, x^{(i)}\right) \right)^2$$

$x^{(i)}$





$y^{(i)}$

$$\begin{bmatrix}0\\0\\0\\0\\0\\1\\0\\0\\0\\0\end{bmatrix} \begin{bmatrix}1\\0\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix} \begin{bmatrix}0\\0\\0\\0\\1\\0\\0\\0\\0\\0\end{bmatrix} \begin{bmatrix}0\\1\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix} \begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\0\\1\end{bmatrix} \begin{bmatrix}0\\0\\1\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}$$

General optimization algorithm

- Specify some initial guess of the solution w(0)
- For k=0,1,…
  - If w(k) is optimal, stop
  - Determine an improved estimate of the solution

$$w(k+1) = w(k) + \alpha(k) * p(k)$$

$\alpha(k)$: a scalar for step length
$p(k)$: a search direction

- The simplest method with a descent direction defined as

$$p(k) = -\nabla f\big(w(k)\big)$$

- For objective written as a sum of large N terms, SGD is typically used

$$\nabla f(w) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(w) \approx \frac{1}{N_m}\sum_{i=1}^{N_m} \nabla f_i(w)$$

An example of writing down the neuron network model

- A simple neuron network with C as our objective (cost) function

$$\begin{tikzcd}\text{Input} & t \arrow[r, "w_1"] & b_1 \arrow[r, "w_2"] & b_2 \arrow[r, "w_3"] & b_3 \arrow[r, "w_4"] & b_4 \arrow[r] & C\end{tikzcd}$$

$C = C(o_4)$

$o_4 = g(z_4) = g(b_4 + w_4 o_3)$

$o_3 = g(z_3) = g(b_3 + w_3 o_2)$

Activation function: $g(z)$

$o_2 = g(z_2) = g(b_2 + w_2 o_1)$

$o_1 = g(z_1) = g(b_1 + w_1 t)$

$$C = C(o_4)$$
$$o_4 = g(z_4) = g(b_4 + w_4 o_3)$$
$$o_3 = g(z_3) = g(b_3 + w_3 o_2)$$
$$o_2 = g(z_2) = g(b_2 + w_2 o_1)$$
$$o_1 = g(z_1) = g(b_1 + w_1 t)$$

$$C = C(g(b_4 + w_4 g(b_3 + w_3 g(b_2 + w_2 g(b_1 + w_1 t)))))$$

$$C = h(w_1, b_1, w_2, b_2, w_3, b_3, w_4, b_4)$$

An example to derive the gradient

Make use of chain-rule

$$C = h(w_1, b_1, w_2, b_2, w_3, b_3, w_4, b_4)$$

$$\begin{tikzcd} t \arrow[r, "w_1"] & b_1 \arrow[r, "w_2"] & b_2 \arrow[r, "w_3"] & b_3 \arrow[r, "w_4"] & b_4 \arrow[r] & C \end{tikzcd}$$

$$C = C(o_4)$$
$$o_4 = g(z_4) = g(b_4 + w_4 o_3)$$
$$o_3 = g(z_3) = g(b_3 + w_3 o_2)$$
$$o_2 = g(z_2) = g(b_2 + w_2 o_1)$$
$$o_1 = g(z_1) = g(b_1 + w_1 t)$$

$$\frac{\partial C}{\partial b_4} = \frac{\partial z_4}{\partial b_4}\frac{\partial o_4}{\partial z_4}\frac{\partial C}{\partial o_4} = g'(z_4)\cdot\frac{\partial C}{\partial o_4}$$

$$\frac{\partial C}{\partial w_4} = \frac{\partial z_4}{\partial w_4}\frac{\partial o_4}{\partial z_4}\frac{\partial C}{\partial o_4} = g(z_3)g'(z_4)\cdot\frac{\partial C}{\partial o_4}$$

$$\frac{\partial C}{\partial b_3} = \frac{\partial z_3}{\partial b_3}\frac{\partial o_3}{\partial z_3}\frac{\partial z_4}{\partial o_3}\frac{\partial o_4}{\partial z_4}\frac{\partial C}{\partial o_4} = g'(z_3)w_4 g'(z_4)\cdot\frac{\partial C}{\partial o_4}$$

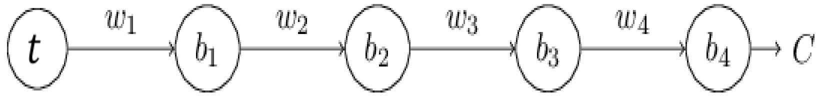
$$\frac{\partial C}{\partial b_2} = \frac{\partial z_2}{\partial b_2}\frac{\partial o_2}{\partial z_2}\frac{\partial z_3}{\partial o_2}\frac{\partial o_3}{\partial z_3}\frac{\partial z_4}{\partial o_3}\frac{\partial o_4}{\partial z_4}\frac{\partial C}{\partial o_4} = g'(z_2)w_3 g'(z_3)w_4 g'(z_4)\cdot\frac{\partial C}{\partial o_4}$$

$$\frac{\partial C}{\partial b_1} = \frac{\partial z_1}{\partial b_1}\frac{\partial o_1}{\partial z_1}\frac{\partial z_2}{\partial o_1}\frac{\partial o_2}{\partial z_2}\frac{\partial z_3}{\partial o_2}\frac{\partial o_3}{\partial z_3}\frac{\partial z_4}{\partial o_3}\frac{\partial o_4}{\partial z_4}\frac{\partial C}{\partial o_4} = g'(z_1)w_2 g'(z_2)w_3 g'(z_3)w_4 g'(z_4)\cdot\frac{\partial C}{\partial o_4}$$

Why is it beneficial to have a network structure?

- A well-organized network structure helps us to “organize” our gradient computation much easier

$$\frac{\partial C}{\partial b_1} = g'(z_1)w_2g'(z_2)w_3g'(z_3)w_4g'(z_4) \cdot \frac{\partial C}{\partial o_4}$$



$$\frac{\partial C}{\partial b_4} = g'(z_4) \cdot \frac{\partial C}{\partial o_4}$$

$$\frac{\partial C}{\partial b_3} = g'(z_3)w_4g'(z_4) \cdot \frac{\partial C}{\partial o_4}$$

$$\frac{\partial C}{\partial b_2} = g'(z_2)w_3g'(z_3)w_4g'(z_4) \cdot \frac{\partial C}{\partial o_4}$$

$$\frac{\partial C}{\partial b_1} = g'(z_1)w_2g'(z_2)w_3g'(z_3)w_4g'(z_4) \cdot \frac{\partial C}{\partial o_4}$$

Backpropagation

- A fancy way to say “gradient computation for neural networks”
- Find out how changing the weights and bias (model parameters) changes the cost function

  We will skip the formal proof here

$$\begin{tikzcd} t \arrow[r, "w_1"] & b_1 \arrow[r, "w_2"] & b_2 \arrow[r, "w_3"] & b_3 \arrow[r, "w_4"] & b_4 \arrow[r] & C \end{tikzcd}$$

$$\frac{\partial C}{\partial b_4} = g'(z_4) \cdot \frac{\partial C}{\partial o_4}$$

$$\frac{\partial C}{\partial b_3} = g'(z_3) w_4 g'(z_4) \cdot \frac{\partial C}{\partial o_4}$$

$$\frac{\partial C}{\partial b_2} = g'(z_2) w_3 g'(z_3) w_4 g'(z_4) \cdot \frac{\partial C}{\partial o_4}$$

$$\frac{\partial C}{\partial b_1} = g'(z_1) w_2 g'(z_2) w_3 g'(z_3) w_4 g'(z_4) \cdot \frac{\partial C}{\partial o_4}$$

Properties of backpropagation

- Requirements for cost functions for backpropagation
  - –Cost can be written as a function of the outputs from the neural network
  - –Cost is average cost over all individual training examples

$$\min_w : f(w) = \frac{1}{2N}\sum_{i=1}^{N}\left(y^{(i)} - h\left(w, x^{(i)}\right)\right)^2$$

- Backpropagation simultaneously compute all the partial derivatives using just one forward pass through the network, followed by one backward pass through the network.
  - –Roughly speaking, the computational cost of the backward pass is about the same as the forward pass
- In other words, the backpropagation algorithm is a clever way of keeping track of small perturbations to the weights (and biases) as they propagate through the network, reach the output, and then affect the cost

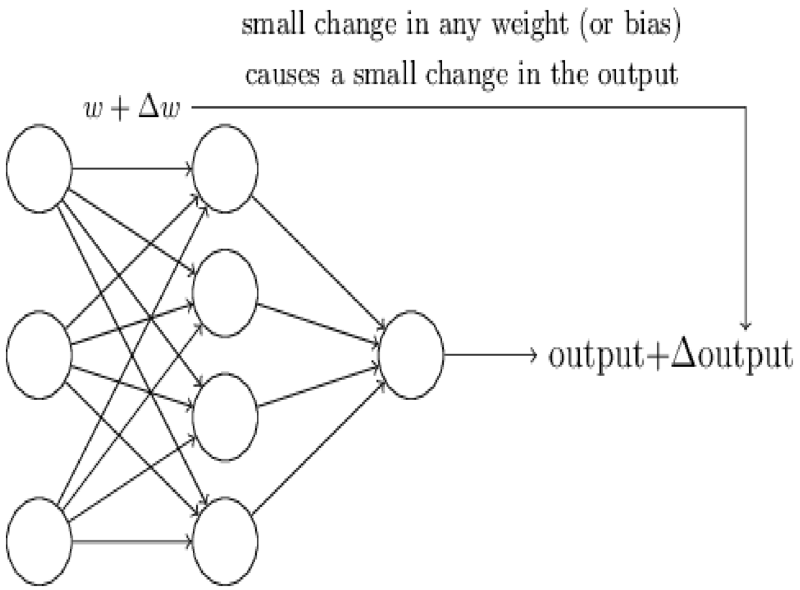
Agenda

- Artificial neuron network for multi-class classification
- Inference: forward propagation
- Training: backpropagation
    - Math
    - A quick review of unconstrained optimization
    - SGD and backpropagation
- Finite-differencing and backpropagation
- Cross-entropy function as objective
- Overfitting issues of neural network
- Regularization techniques to overcome overfitting
- Initialization and softmax output layer
- Conclusions

A finite differencing method for computing gradient

- Recall: the first order derivative of a function at point $(w, f(w))$ is defined as the instantaneous rate of change at that point, which is given by the limit of the average rate of change as

$$\frac{df(w)}{dw} = \lim_{\Delta w \to 0} \frac{f(w + \Delta w) - f(w)}{\Delta w}$$


small change in any weight (or bias)
causes a small change in the output
$w + \Delta w$
output+Δoutput

Comparison between finite-differencing and backpropagation

- Both can compute gradients reasonably well
- Finite differencing is simple, but it's an approximated solution, and it can be way too costly to compute

  For every parameter, you need to perturb the parameter value, and then do a forward propagation to obtain the perturbed output

$$\frac{\partial f(w_i)}{\partial w_i} \approx \frac{f(w_i + \Delta w_i) - f(w_i)}{\Delta w_i}$$

  But it's a good tool to verify the correctness of backpropagation's implementation

- Backpropagation can be accurate, but it's quite complicated to implement and can suffer other stability issues

  Backprogation is typically done for a single training example, and then averaged across all training examples

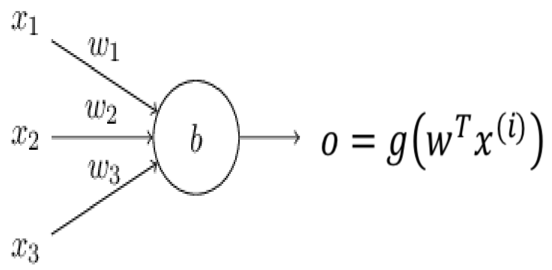
$$\nabla f(w) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(w) \approx \frac{1}{N_m}\sum_{i=1}^{N_m} \nabla f_i(w)$$

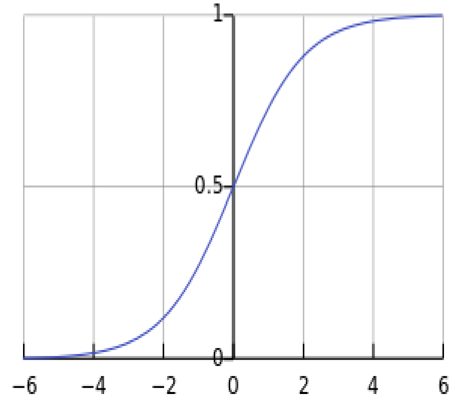
Agenda

- Artificial neuron network for multi-class classification
- Inference: forward propagation
- Training: backpropagation
  - Math
  - A quick review of unconstrained optimization
  - SGD and backpropagation
- Finite-differencing and backpropagation
- Cross-entropy function as objective
- Overfitting issues of neural network
- Regularization techniques to overcome overfitting
- Initialization and softmax output layer
- Conclusions

Cross-entropy cost function for single neuron

- The cost function we used for logistic regression





$$\min_{w}: \ f(w) = \frac{1}{N}\sum_{i=1}^{N}\left[-y^{(i)} log\left(g\left(w^T x^{(i)}\right)\right) - \left(1 - y^{(i)}\right) log\left(1 - g\left(w^T x^{(i)}\right)\right)\right]$$

We derived this cost function based on maximum likelihood

- This cost function is also called cross-entropy function

$$p(k)_j = -\frac{\partial f}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - g\left(w^T x^{(i)}\right)\right) x_j^{(i)}$$

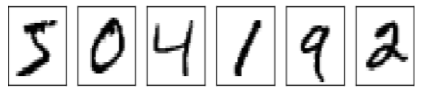
Agenda

- Artificial neuron network for multi-class classification
- Inference: forward propagation
- Training: backpropagation
    - Math
    - A quick review of unconstrained optimization
    - SGD and backpropagation
- Finite-differencing and backpropagation
- Cross-entropy function as objective
- Overfitting issues of neural network
- Regularization techniques to overcome overfitting
- Initialization and softmax output layer
- Conclusions

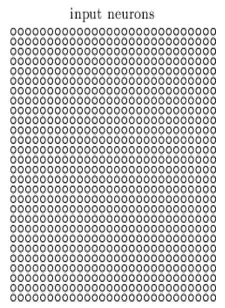
An example

- Neuron network with 30 hidden neurons for MINIST digital recognition

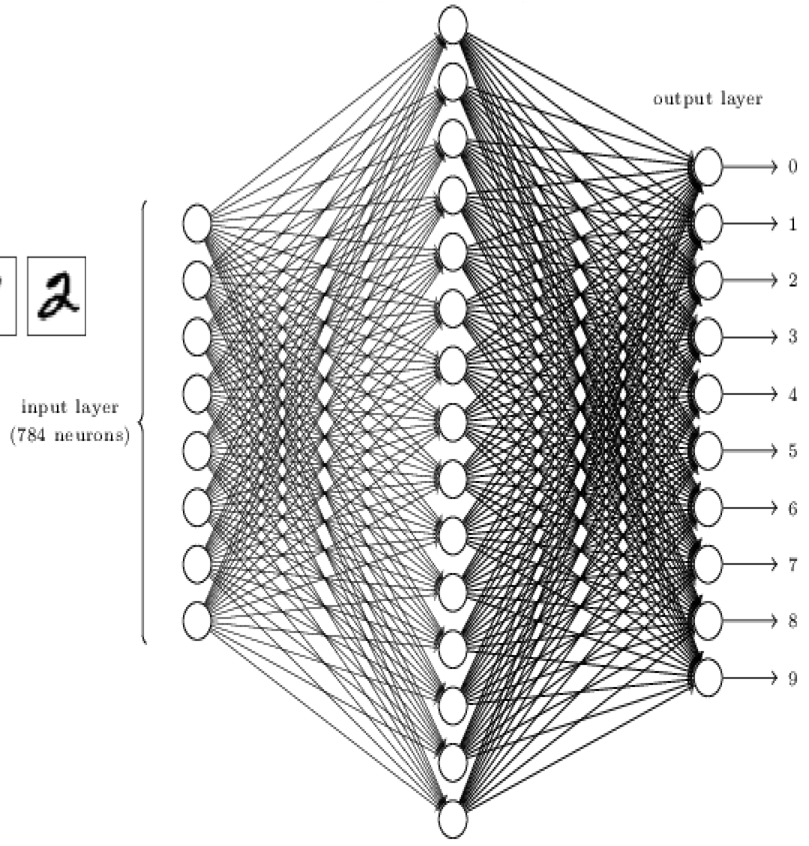
$x^{(i)}$



28x28 pixels



Hidden layer
(n = 30 neurons)



$y^{(i)}$

$$\begin{bmatrix}0\\0\\0\\0\\0\\1\\0\\0\\0\\0\end{bmatrix} \begin{bmatrix}1\\0\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix} \begin{bmatrix}0\\0\\0\\0\\1\\0\\0\\0\\0\\0\end{bmatrix} \begin{bmatrix}0\\1\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix} \begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\0\\1\end{bmatrix} \begin{bmatrix}0\\0\\1\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}$$

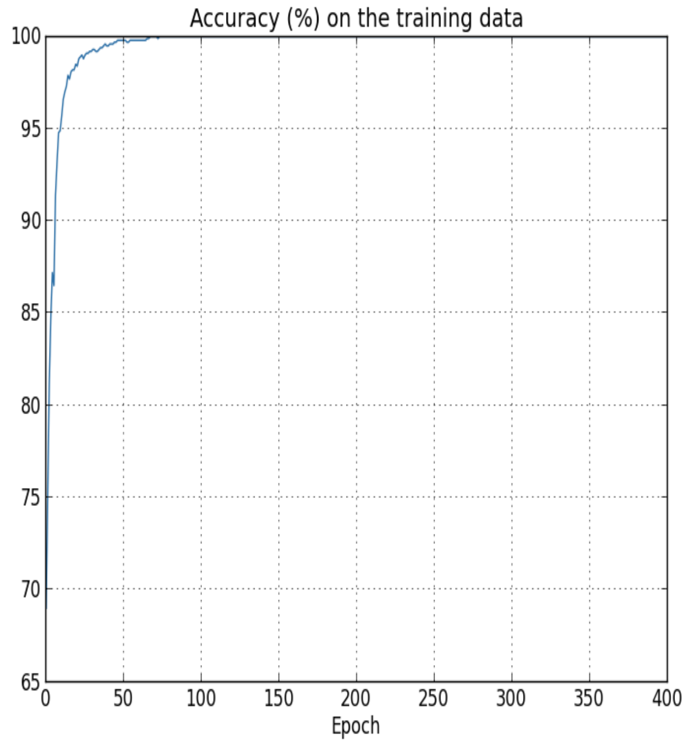
Issues of overfitting

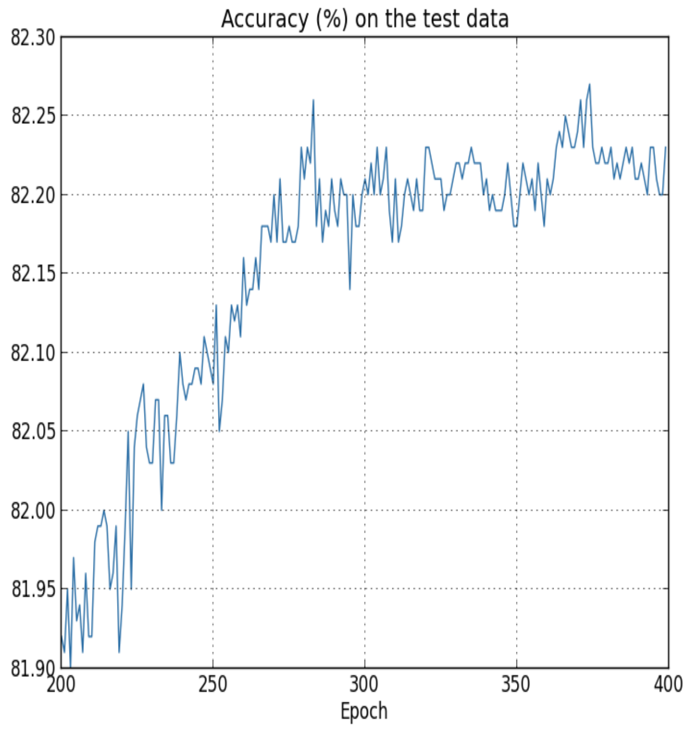
- Number of model parameters

  (28*28)*30+30+30*10+10 = 23,860

- Train the network with 1,000 training images

  The # of parameters >> the # of training data





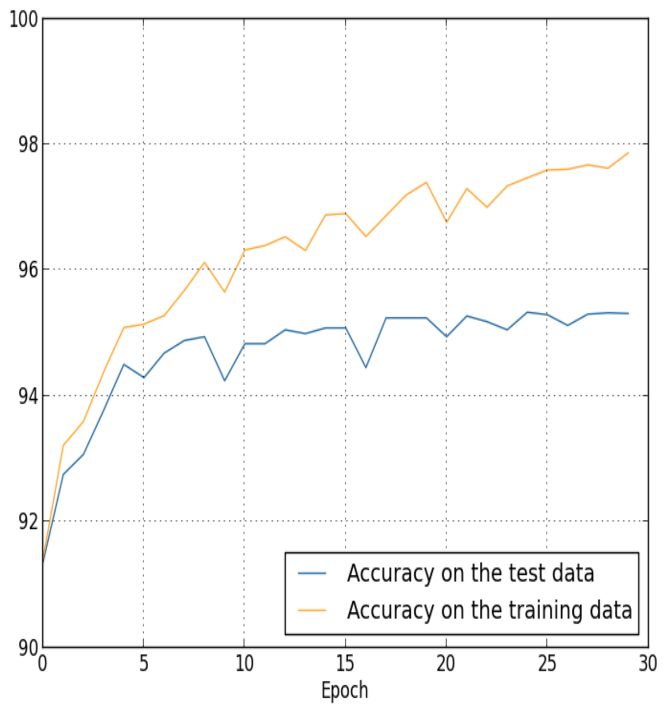
Issues of overfitting

- Number of model parameters

  (28*28)*30+30+30*10+10 = 23,860

- Train the network with 50,000 training images

  The # of parameters < the # of training data



Agenda

- Artificial neuron network for multi-class classification
- Inference: forward propagation
- Training: backpropagation
    - Math
    - A quick review of unconstrained optimization
    - SGD and backpropagation
- Finite-differencing and backpropagation
- Cross-entropy function as objective
- Overfitting issues of neural network
- Regularization techniques to overcome overfitting
- Initialization and softmax output layer
- Conclusions

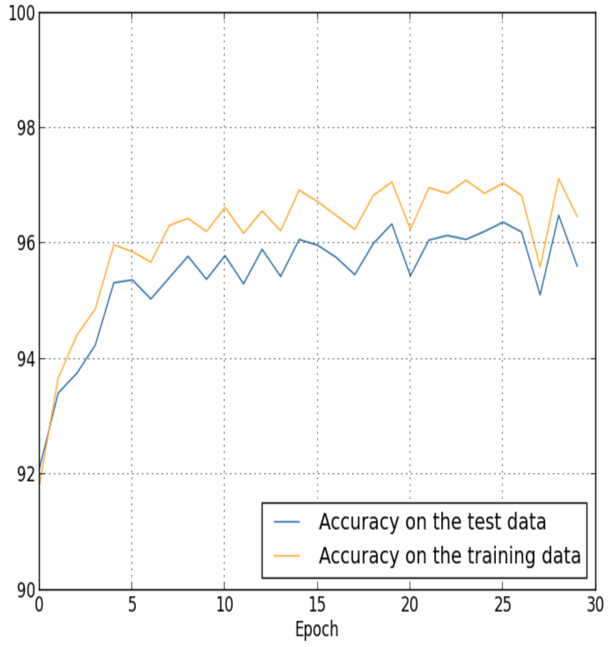
Overfitting and regularization

- Overfitting is a major problem in neural networks

  This is especially true in modern networks, which often have very large numbers of weights and biases

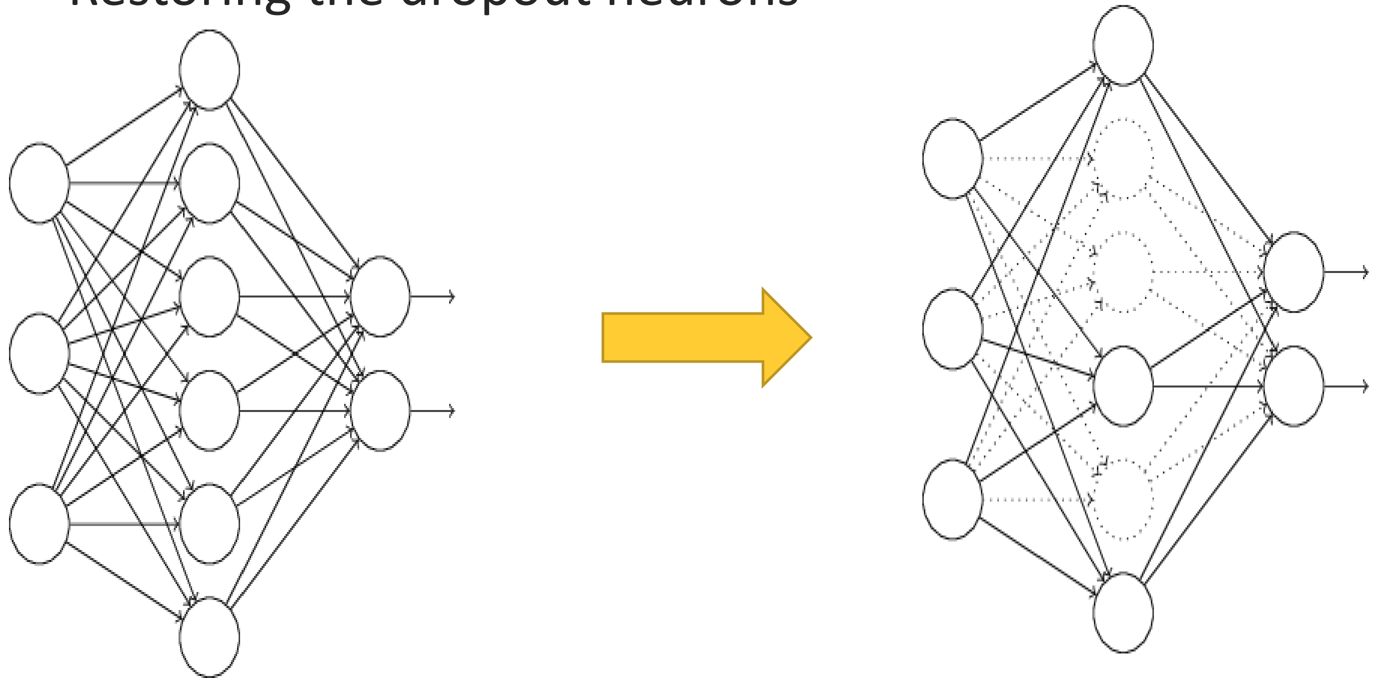
- Regularization helps to overcome overfitting

$$\min_{w}: \; f(w) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\left[-y^{(i)}log\left(g_j\left(w^T x^{(i)}\right)\right) - \left(1-y^{(i)}\right)log\left(1-g_j\left(w^T x^{(i)}\right)\right)\right] + \frac{\lambda}{2N}\sum_{k} w_k^2$$



Dropout as another regularization technique for neural networks

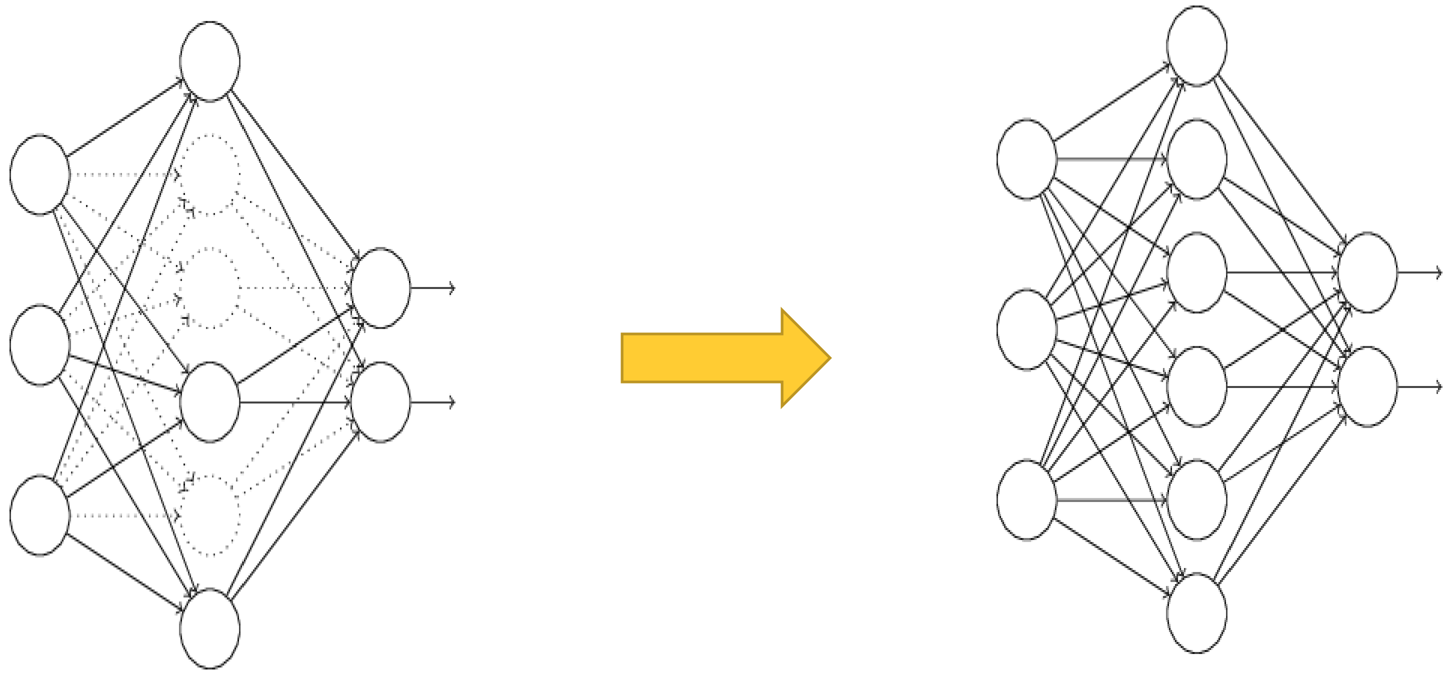
- Dropout does not modify the cost function, but modify the network itself
  - –For each iteration over a mini-batch
    - ▪We randomly (but temporarily) delete or ignore half the hidden neurons in the network, while leaving the input and output neurons untouched
    - ▪Forward-propagate the input through the modified network, and then backpropagate the result, also through the modified network
    - ▪Update the appropriate weights and biases.
    - ▪Restoring the dropout neurons



Dropout as another regularization technique for neural networks

- When we actually run the full network, twice as many hidden neurons will be active

  To compensate for that, we halve the weights outgoing from the hidden neurons



- Intuitively, why dropout may work?

  It' s like training different networks (with reduced parameters) and then take an averaging from them

Artificially expanding the training data as another regularization technique

- One reason for overfitting is the lack of training data → so why not creating more training data
  - –Obtaining labeled data is more difficult (too costly)
  - –So slightly modifying existing labeled data to create new data with the same label

The original MNIST digits

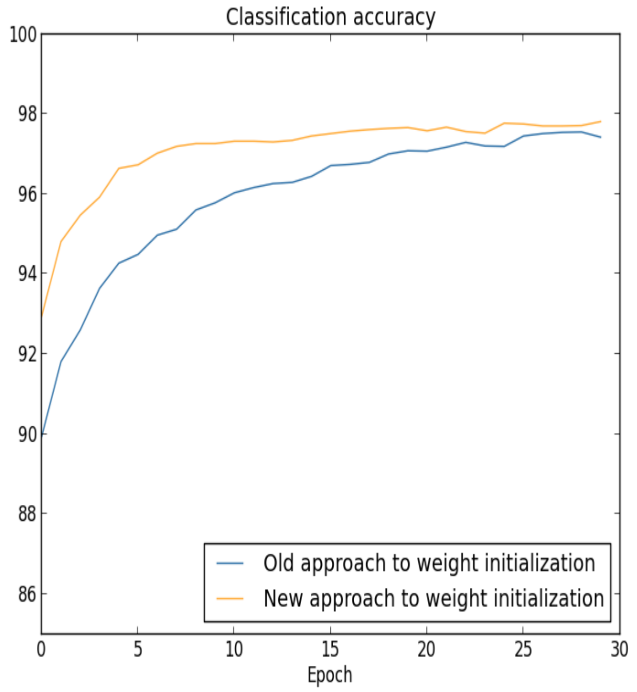
[http://www.cs.toronto.edu/~tijmen/affNIST/]

Agenda

- Artificial neuron network for multi-class classification
- Inference: forward propagation
- Training: backpropagation
  - Math
  - A quick review of unconstrained optimization
  - SGD and backpropagation
- Finite-differencing and backpropagation
- Cross-entropy function as objective
- Overfitting issues of neural network
- Regularization techniques to overcome overfitting
- Initialization and softmax output layer
- Conclusions

Parameter initialization

- Straightforward strategy: randomly generate values based on independent Gaussian random variables, normalized to have mean 0 and standard deviation 1
- It turns out different strategies for initialization can lead to different results with different quality



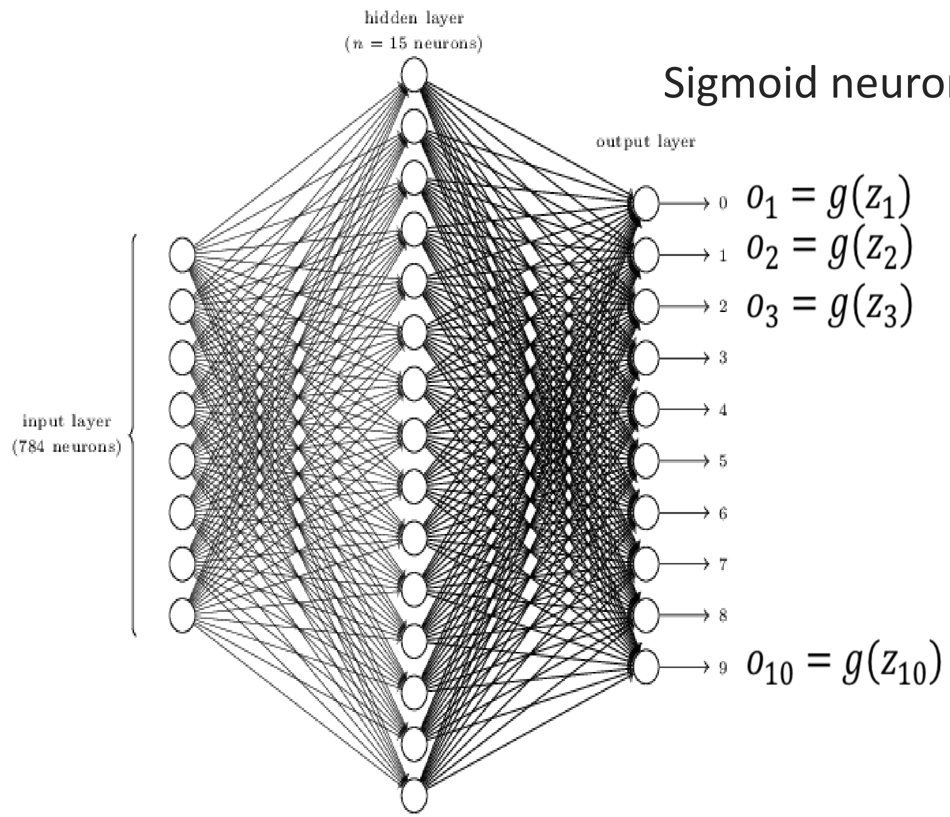
- How to best define initialization strategy is still an open research problem

Softmax layer to transform outputs into a probability distribution

- For multi-class neural network, the output class is determined by the maximum value among all output neurons (after the forward propagation)

- Softmax layer transform the outputs into a probability distribution
  - –The one with the highest probability determines the class

Sigmoid neuron output



$o_1 = g(z_1)$

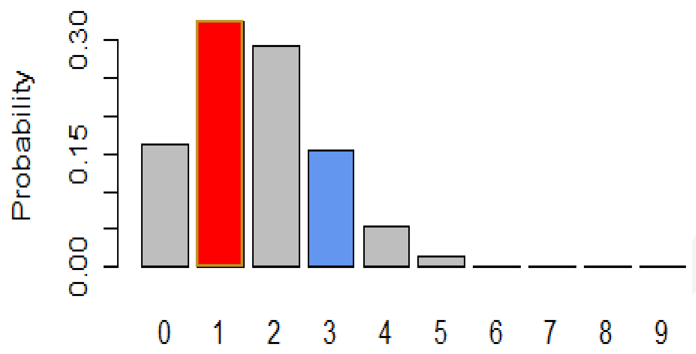
$o_2 = g(z_2)$

$o_3 = g(z_3)$

$o_{10} = g(z_{10})$

Softmax neuron output

$$o_{j,new} = \frac{e^{z_j}}{\sum_{j=1}^{K} e^{z_j}}$$



Agenda

- Artificial neuron network for multi-class classification
- Inference: forward propagation
- Training: backpropagation
  - Math
  - A quick review of unconstrained optimization
  - SGD and backpropagation
- Finite-differencing and backpropagation
- Cross-entropy function as objective
- Overfitting issues of neural network
- Regularization techniques to overcome overfitting
- Initialization and softmax output layer
- Conclusions

Conclusions

- Artificial neural network is a machine learning model (in part) inspired by the neuroscience's understanding of biological neurons
  Multi-class classification can be easily formulated as a ANN problem

- Forward propagation and backpropagation in SGD are key algorithms for ANN

- Overfitting is a major issues for ANN training, and a number of regularization techniques are developed to address this issue

- ANN has been around for many years, and it only recently gained popularity

Lab 2: Learning XOR with MLP on Google Colab

Assignments and Related Documents:

- https://jqub.github.io/2021/09/01/ML4Emb/

Due Date: This Friday (xx-xx-xxxx) by 1 PM

GEORGE MASON UNIVERSITY

GMU.EDU

George Mason University

4400 University Drive
Fairfax, Virginia 22030

Tel: (703)993-1000