Neural Network Introduction

LING 570

Fei Xia

(Some slides are based on Geoffrey Hinton’s lectures)

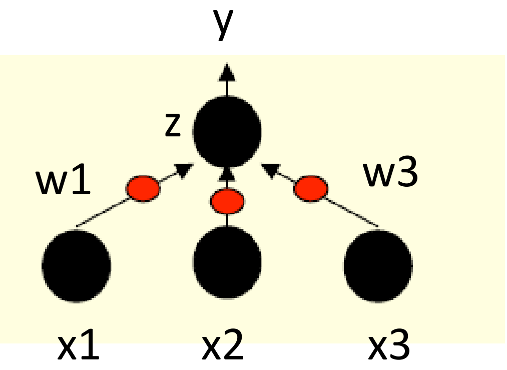
Outline

- What is neural network?

- Learning of linear neurons

- Learning of logistic neurons **

Neurons in human brain



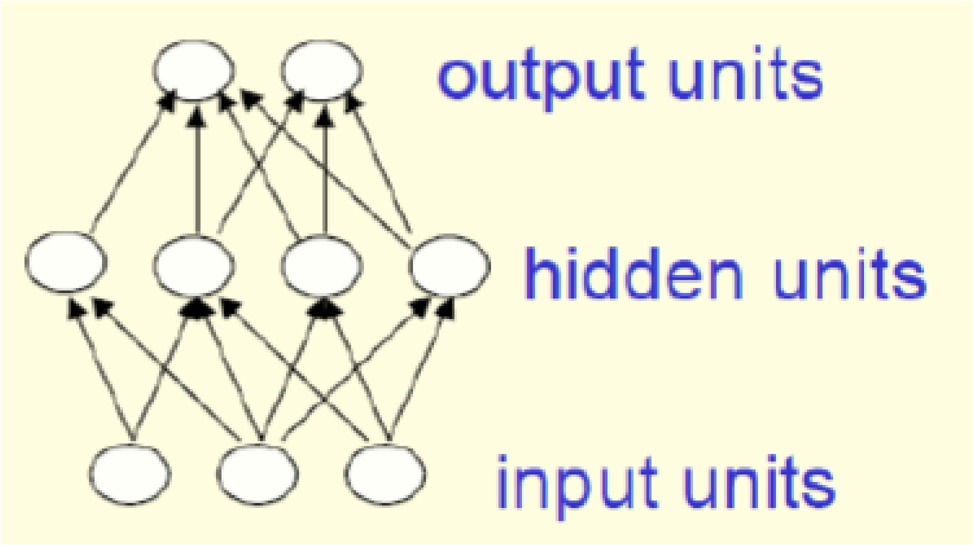
- Each neuron receives inputs from other neurons.
- The effect of each input line on the neuron is controlled by a weight.
- The weights adapt so that the whole network learns to perform useful computations.
- Human brain has about 100 billion neurons, each with about 10K weights.

What is neural network (NN)?

- A graph with multiple layers:
  - input layer
  - output layer
  - zero or more hidden layers

- Each layer has multiple nodes (aka neurons or units)

- Each arc has a weight.



Neural network (NN)

- A modern NN is a network of small computing units.
- Each unit (a neuron) takes a vector of input values and produce a single output.
- The use of NN is often called deep learning, because modern networks are often deep (having many layers).
- It’s best to think of NN as function approximation machines, occasionally drawing some insights from what we know about the brain, rather than as models of brain function.
- Common NN models:
  - Feedforward networks
  - Recurrent NN (RNN)
  - Recursive NN
  - Encoder-decoder model
  - Convolutional NN (CNN)
  - …

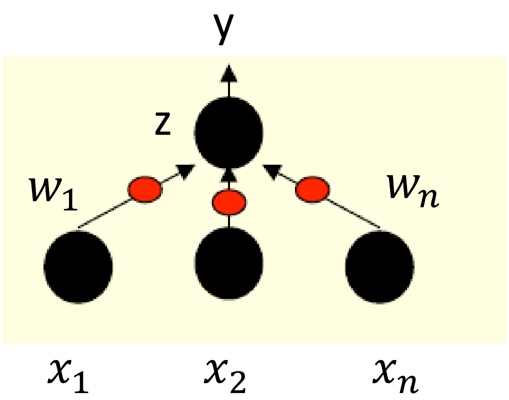
Why now?

- NN is an old algorithm and it was not successful in NLP before 2000s.

- What has changed?
  - New methods for unsupervised pre-training have been developed.
  - More efficient parameter estimation methods
  - Better understanding of model regularization
  - More data; more and faster machines.

➔ NN has produced state-of-the-art results on many NLP tasks.

A neuron



$$z = b + \sum_i x_i \, w_i$$

$b \text{ is called the bias}$

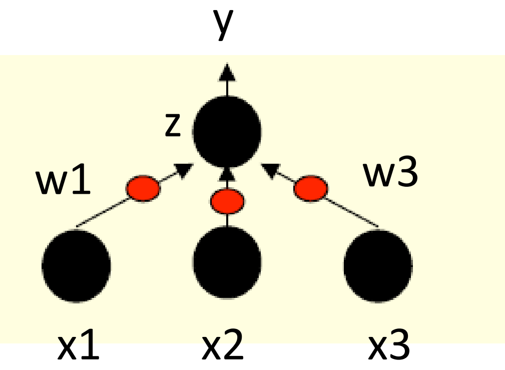
$$y = f(z)$$

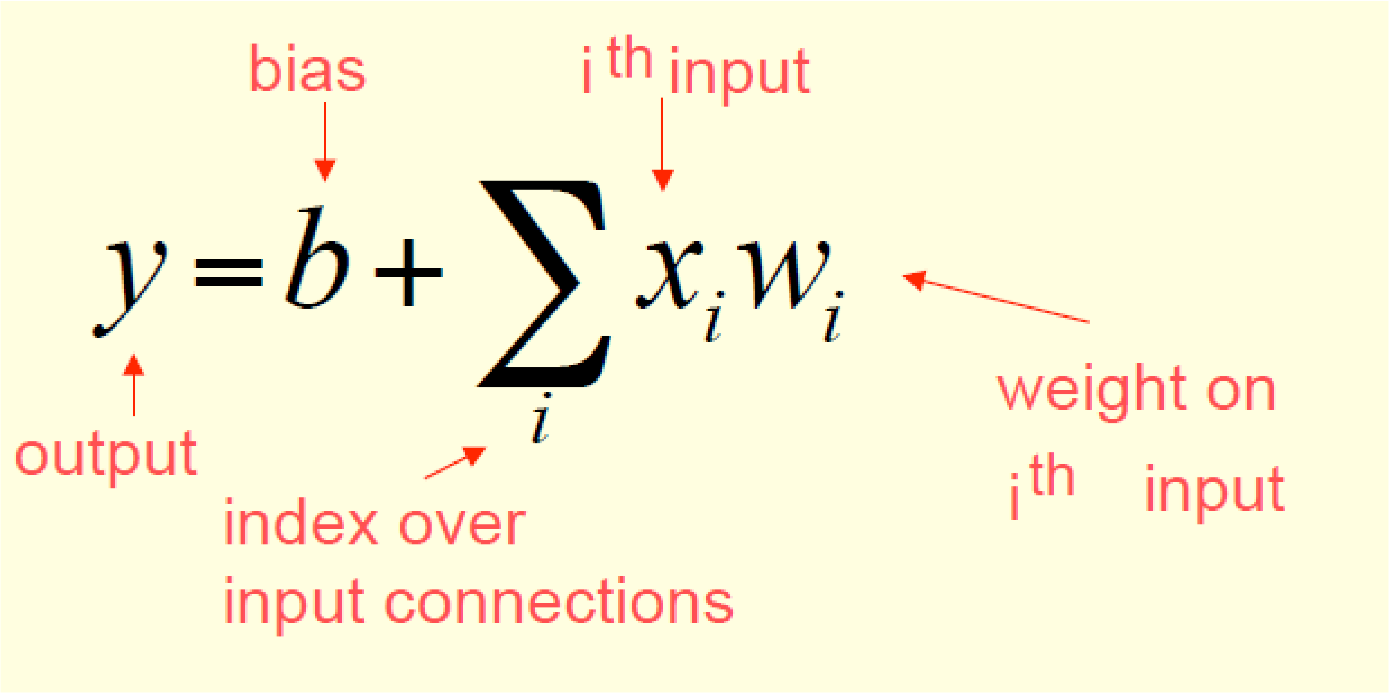
$f \text{ is called an activation function}$

What activation function should we use?

Linear neurons

- The activation function is linear.
- They are simple but computationally limited.





bias

i^{th} input

$$y = b + \sum_i x_i w_i$$

output

index over input connections

weight on i^{th} input

Binary threshold neuron

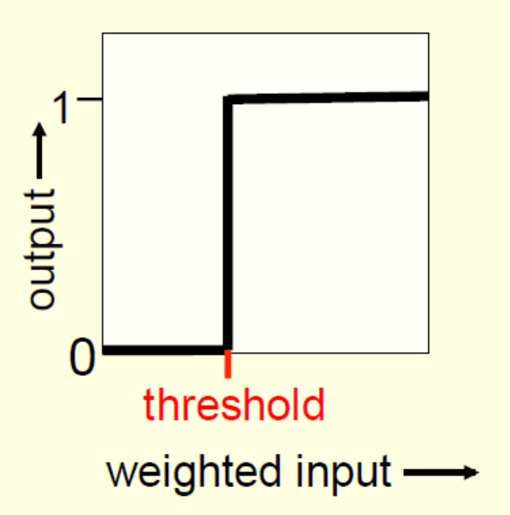
$$z = \sum_i x_i w_i$$

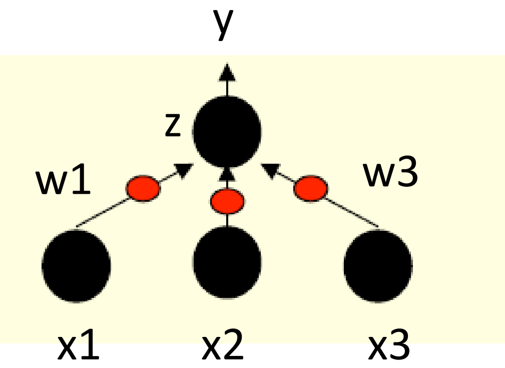
$$y = \begin{cases} 1 \text{ if } z \geq \theta \\ 0 \text{ otherwise} \end{cases}$$

$$\Updownarrow \quad \theta = -b$$

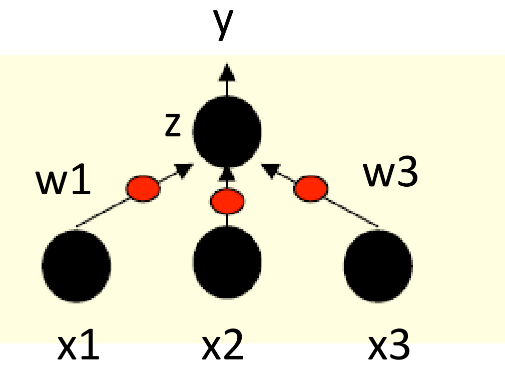
$$z = b + \sum_i x_i w_i$$

$$y = \begin{cases} 1 \text{ if } z \geq 0 \\ 0 \text{ otherwise} \end{cases}$$



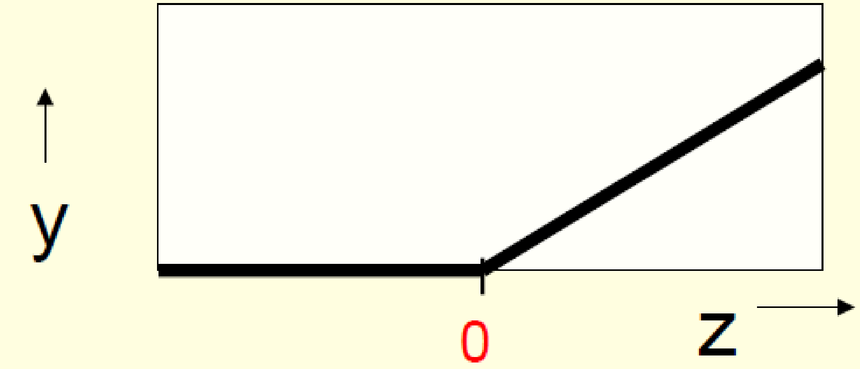


Rectified linear neurons



$$z = b + \sum_i x_i w_i$$

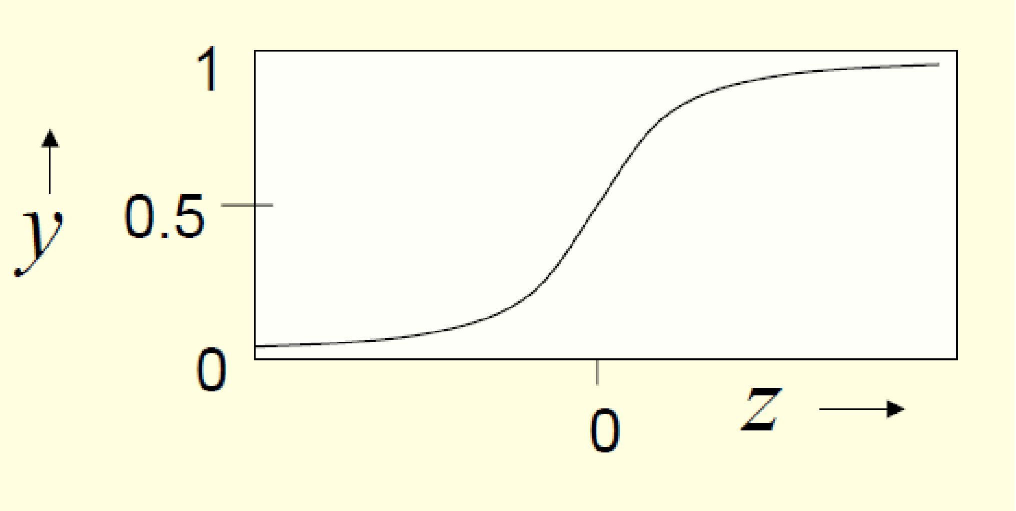
$$y = \begin{cases} z & \text{if } z > 0 \\ 0 & \text{otherwise} \end{cases}$$



Sigmoid neurons

- They give a real-valued output that is a smooth and bounded function of their total input.

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



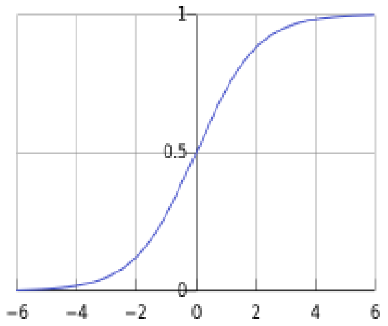
Good derivative property:

$$f'(z) = f(z)(1 - f(z))$$

sigmoid vs. tanh

logistic (“sigmoid”)

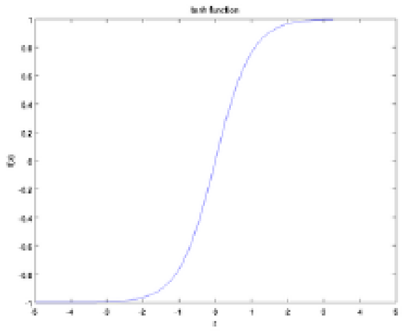
$$f(z) = \frac{1}{1+\exp(-z)}.$$



$$f'(z) = f(z)(1 - f(z))$$

tanh

$$f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}},$$



$$f'(z) = 1 - f(z)^2$$

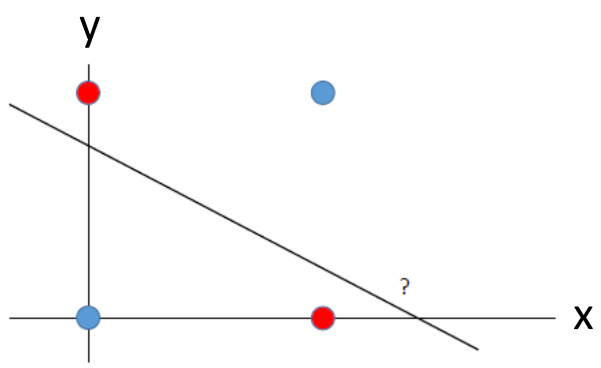
tanh is just a rescaled and shifted sigmoid $\tanh(z) = 2\mathrm{logistic}(2z) - 1$

tanh is what is most used and often performs best for deep nets

Activation function

- There are many activation functions.

- Which one to choose?
  - non-linearity is crucial
  - prefer ones with nice derivatives which makes learning easy.



Let z = x XOR y:
    if x == y,   then z is 0   (see blue circles)
                 else z is 1   (see red circles)

There is no linear classifier that can solve the XOR problem.
  i.e., there is no line that separates blue and red circles.

Hidden layer

- Single neuron: $y = f(b + \sum_i w_i x_i)$

- 1-hidden layer:

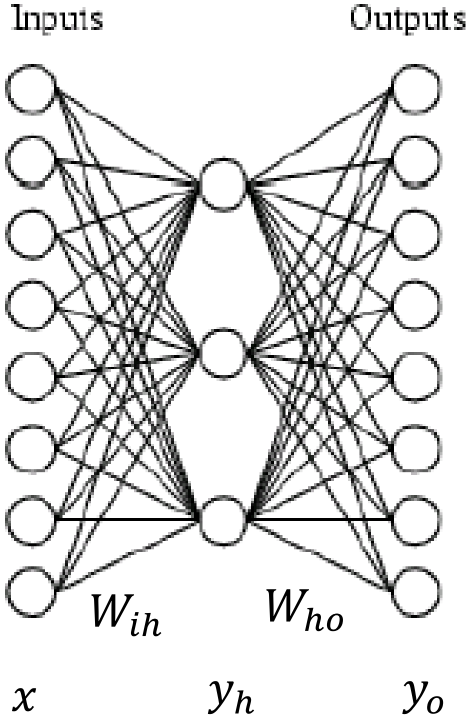
The hidden layer output:

$$y_h = f(W_{ih}\, x)$$

The output layer output:

$$y_o = \ f(W_{ho} y_h)$$

$$= f(W_{ho}\, f(W_{ih} x))$$



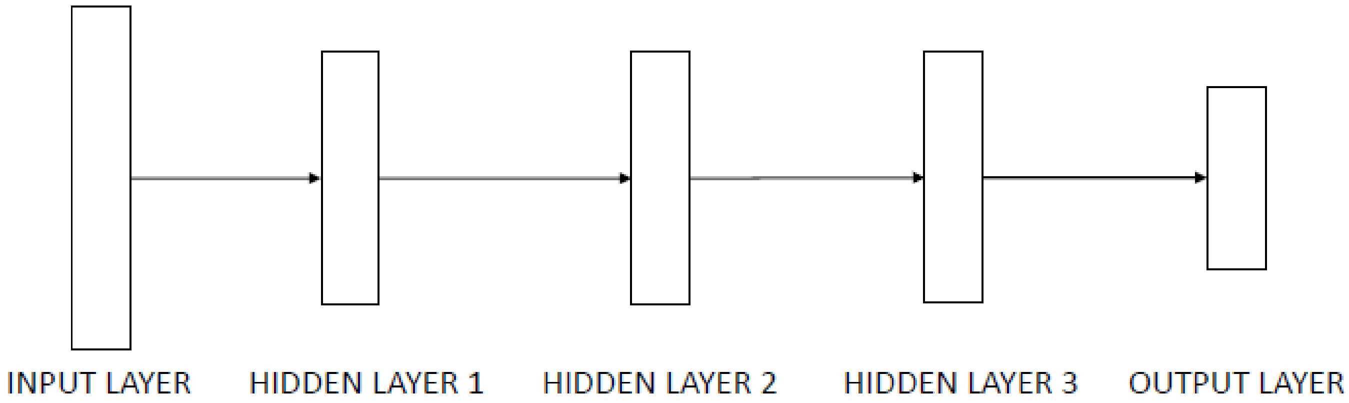
Notation:

if $z = [z_1, z_2, \dots, z_n]$

then $f(z) = [f(z_1), \dots, f(z_n)]$

Deep learning

- Deep nets have many hidden layers in the model.
- Deep learning aims to learn patterns that cannot be learned efficiently with shallow models.
- When we try to learn complex function that is a composition of simpler functions, it may be beneficial to use deep nets.
- Many proposed deep models do not learn anything more than what a shallow model can learn. Beware the hype.



softmax function

- Softmax function turns a vector of arbitrary real values into a probability distribution (i.e., the values are in [0, 1] and sum to one)

- Let $z = [z_1, z_2, \dots, z_n]$
  softmax(z) = $[y_1, y_2, \dots, y_n]$,
  where $y_j = e^{z_j} / \sum_k e^{z_k}$

- In MaxEnt, $z = w\, f(x, y); that\ is, z_j = w\, f\left(x, c_j\right) = \sum_i w_i f_i(x, c_j)$

- In NN, the softmax function is often used as the final layer of an NN for classification.

- For more detail, see https://en.wikipedia.org/wiki/Softmax_function

Why “representation learning”?

- MaxEnt (multinomial logistic regression):

$$y = \mathrm{softmax}(w \cdot \underline{f(x, y)})$$

You design the feature vector

- NNs:

$$y = \mathrm{softmax}(w \cdot \underline{\sigma(Ux)})$$

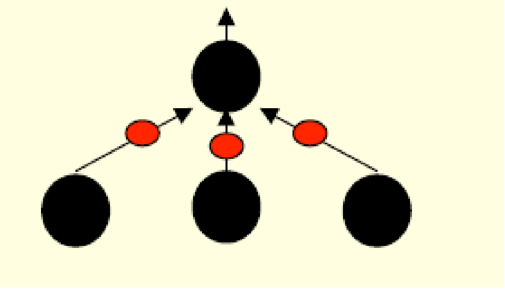
$$y = \mathrm{softmax}(w \cdot \underline{\sigma(U^{(n)}(\ldots\sigma(U^{(2)}\sigma(U^{(1)}x))))})$$

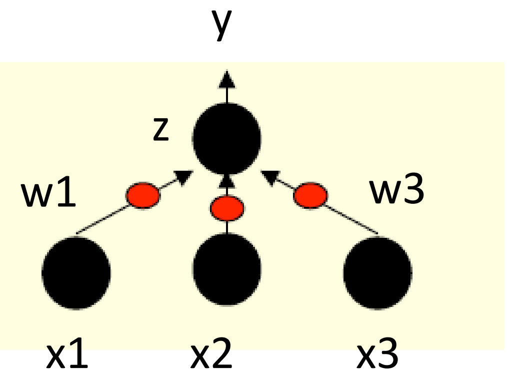
Feature representations are “learned” through hidden layers

Outline

- Neural Network
- Learning of linear neurons
- Learning of logistic neurons **

Learn the weights of linear neuron





bias

i^{th} input

$$y = b + \sum_i x_i w_i$$

output

index over input connections

weight on i^{th} input

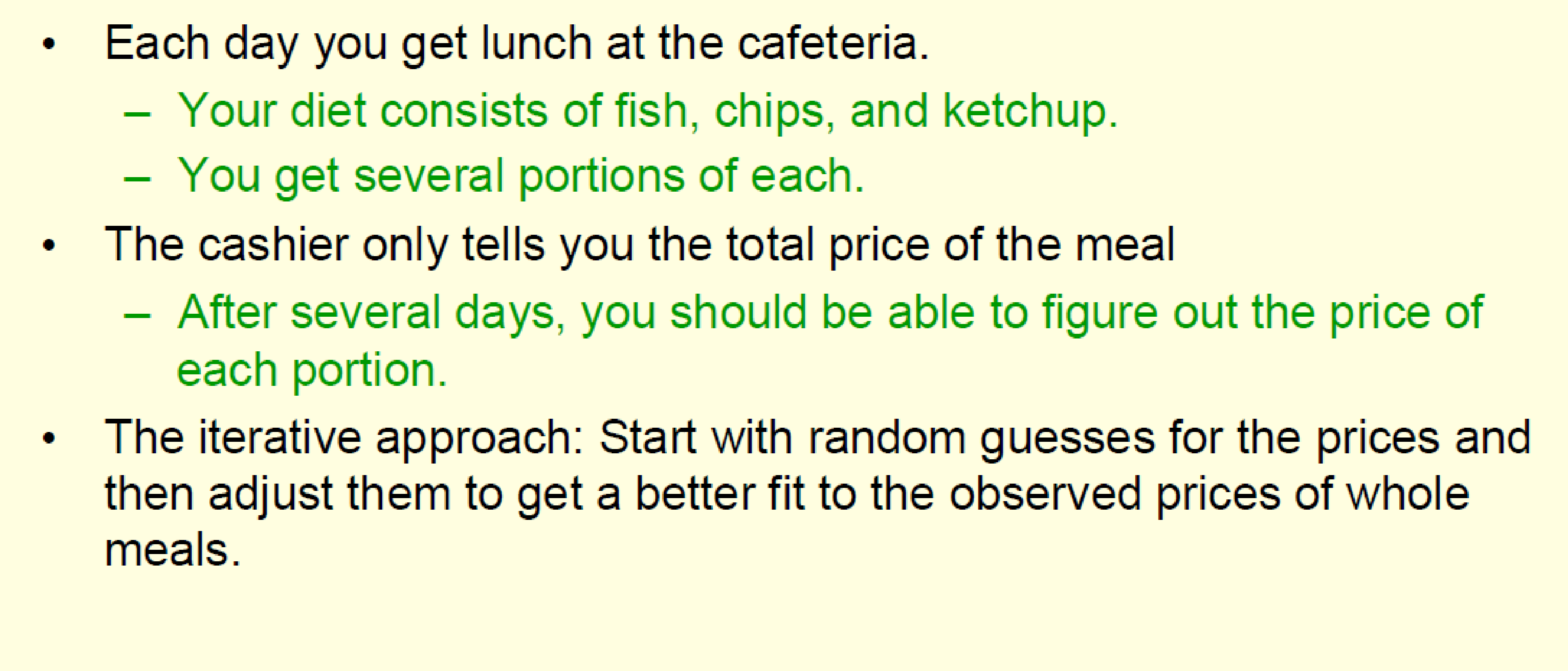
weight vector

$$y = \sum_i w_i x_i = \mathbf{w}^T \mathbf{x}$$

neuron's estimate of the desired output

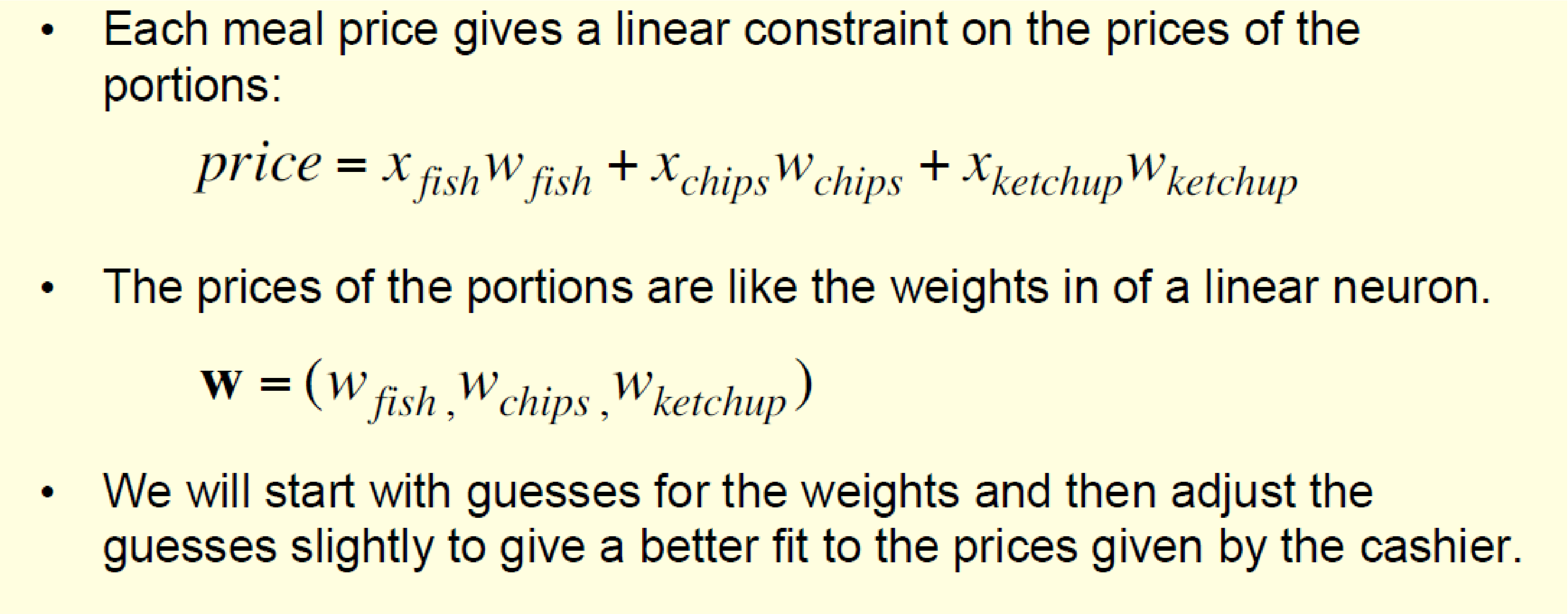
input vector

A toy example



- Each day you get lunch at the cafeteria.
  - Your diet consists of fish, chips, and ketchup.
  - You get several portions of each.
- The cashier only tells you the total price of the meal
  - After several days, you should be able to figure out the price of each portion.
- The iterative approach: Start with random guesses for the prices and then adjust them to get a better fit to the observed prices of whole meals.

Solving the equations iteratively



- Each meal price gives a linear constraint on the prices of the portions:

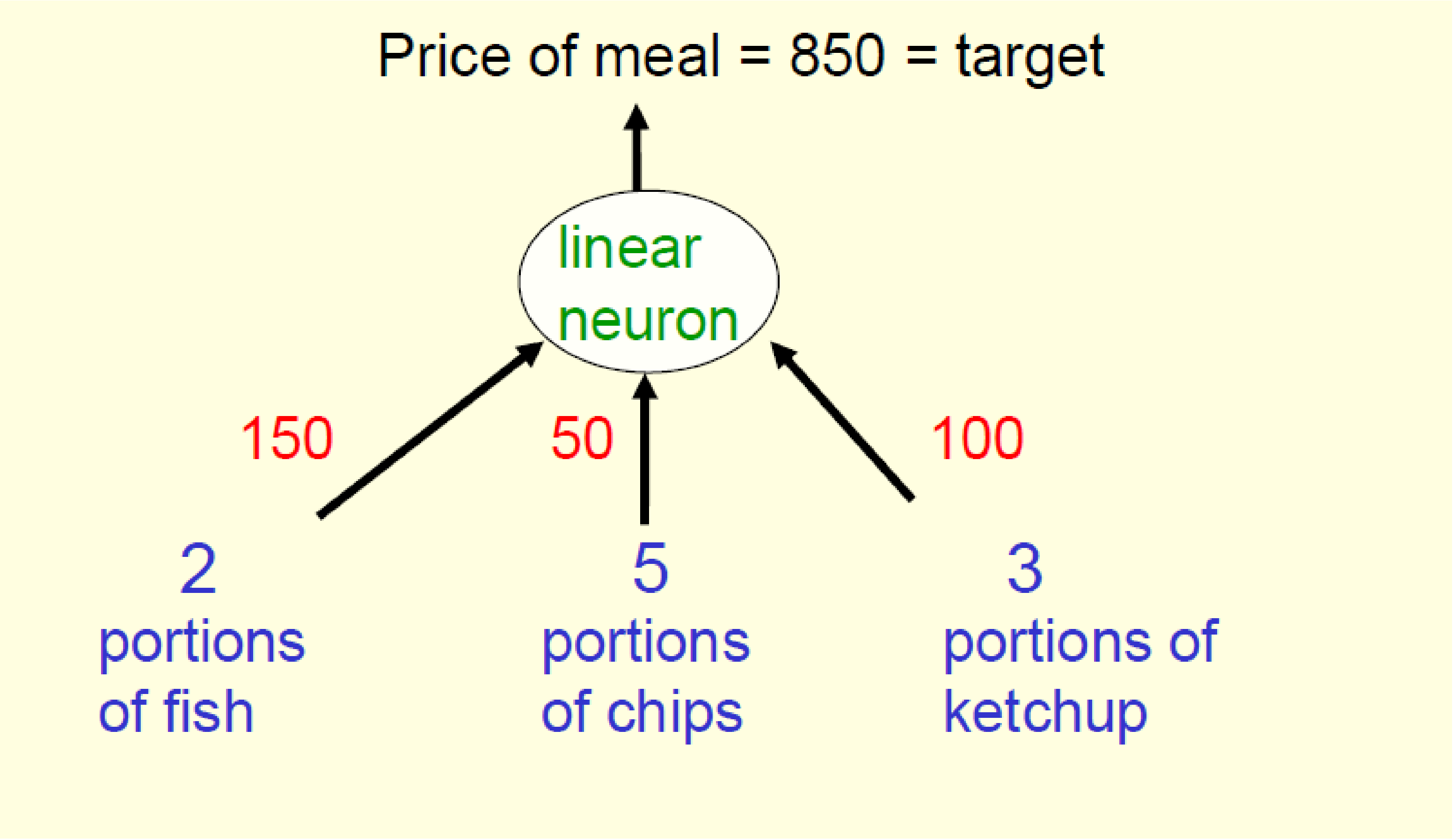
$$price = x_{fish} w_{fish} + x_{chips} w_{chips} + x_{ketchup} w_{ketchup}$$

- The prices of the portions are like the weights in of a linear neuron.

$$\mathbf{w} = (w_{fish}, w_{chips}, w_{ketchup})$$

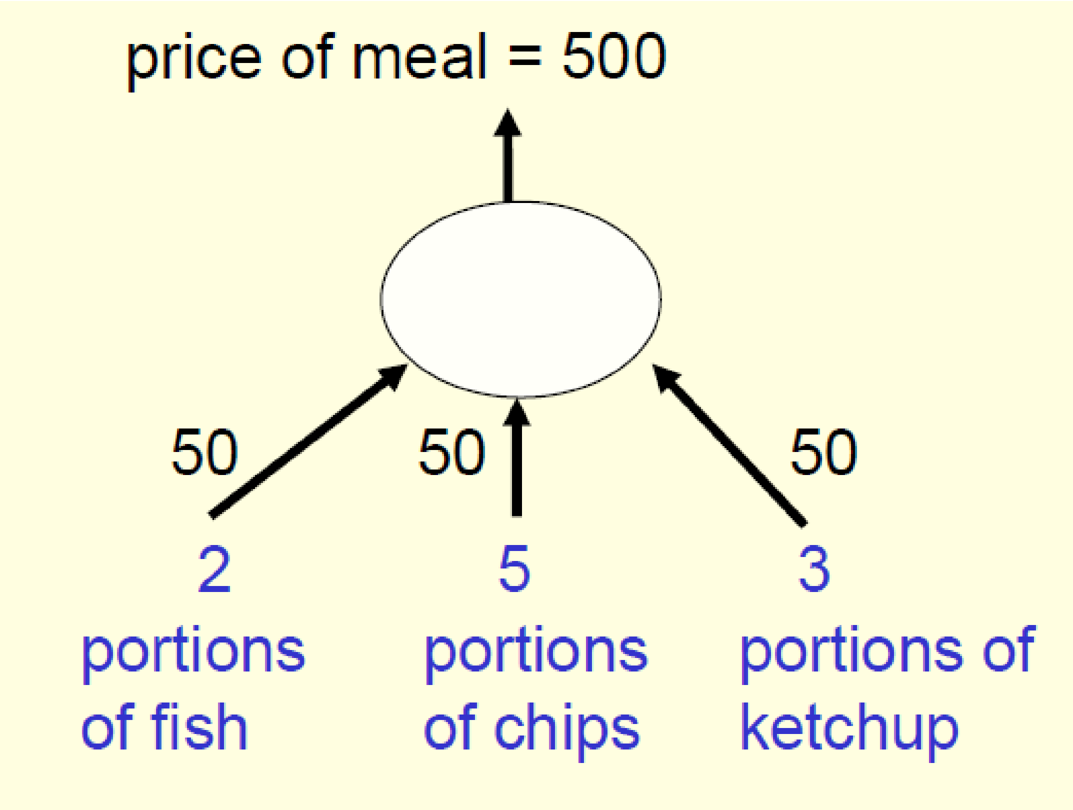
- We will start with guesses for the weights and then adjust the guesses slightly to give a better fit to the prices given by the cashier.

The true weights used by the cashier



Price of meal = 850 = target

Start with arbitrary initial weights

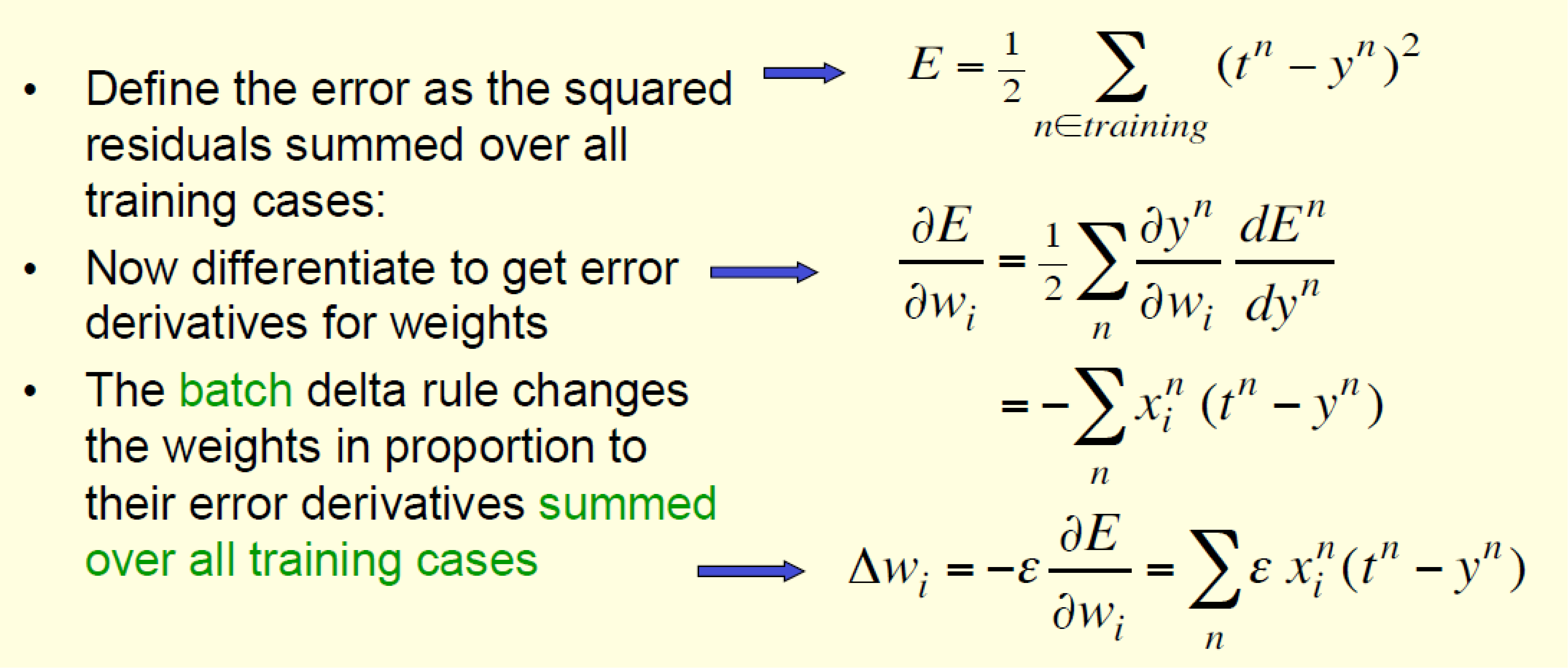


- Residual error = 350
- The “delta-rule” for learning is:
  $$\Delta w_i = \varepsilon \, x_i \, (t - y)$$
- With a learning rate ε of 1/35, the weight changes are +20, +50, +30
- This gives new weights of 70, 100, 80.
  - Notice that the weight for chips got worse!

Truth: (150, 50, 100)

Init value: (50, 50, 50), error=350
➔ (70, 100, 80), error=30
➔ …
➔ … ## Repeat until the residual error is small enough

Deriving the delta rule



- Define the error as the squared residuals summed over all training cases:

$$E = \frac{1}{2} \sum_{n \in training} (t^n - y^n)^2$$

- Now differentiate to get error derivatives for weights

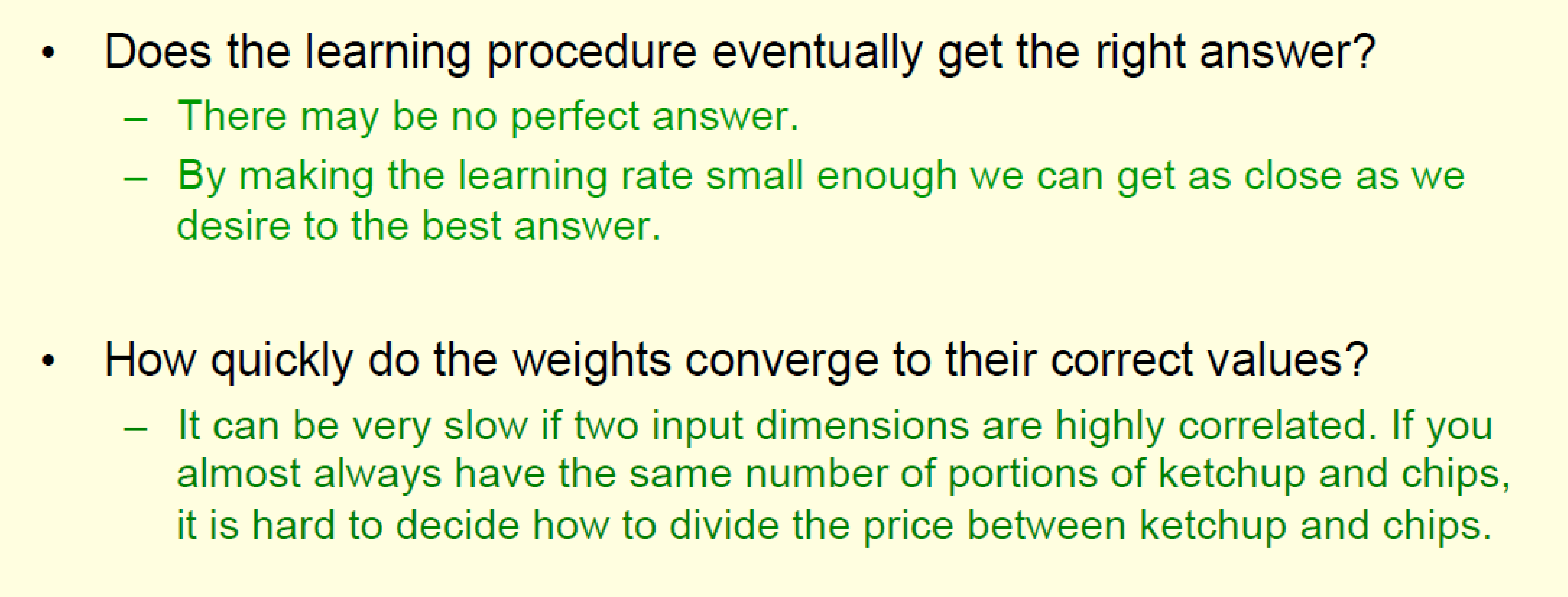
$$\frac{\partial E}{\partial w_i} = \frac{1}{2} \sum_n \frac{\partial y^n}{\partial w_i} \frac{dE^n}{dy^n}$$

$$= -\sum_n x_i^n \, (t^n - y^n)$$

- The batch delta rule changes the weights in proportion to their error derivatives summed over all training cases

$$\Delta w_i = -\varepsilon \frac{\partial E}{\partial w_i} = \sum_n \varepsilon \, x_i^n (t^n - y^n)$$

Behavior of the iterative learning procedure



- Does the learning procedure eventually get the right answer?
  - There may be no perfect answer.
  - By making the learning rate small enough we can get as close as we desire to the best answer.

- How quickly do the weights converge to their correct values?
  - It can be very slow if two input dimensions are highly correlated. If you almost always have the same number of portions of ketchup and chips, it is hard to decide how to divide the price between ketchup and chips.

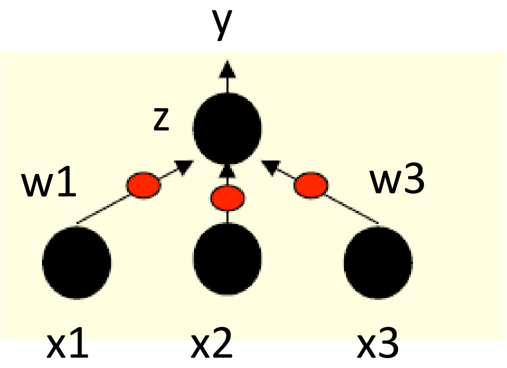
Outline

- What is Neural Network?

- Learning of linear neurons

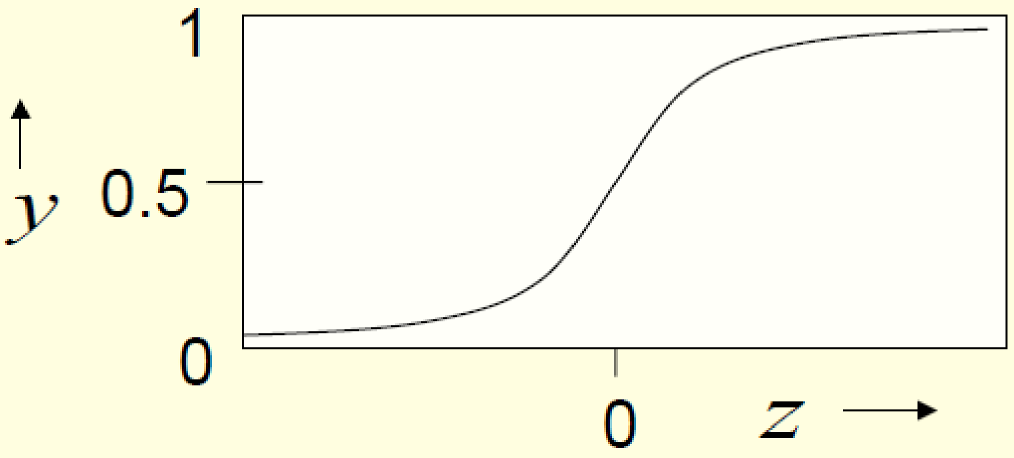
- Learning of logistic neurons **

Logistic neurons

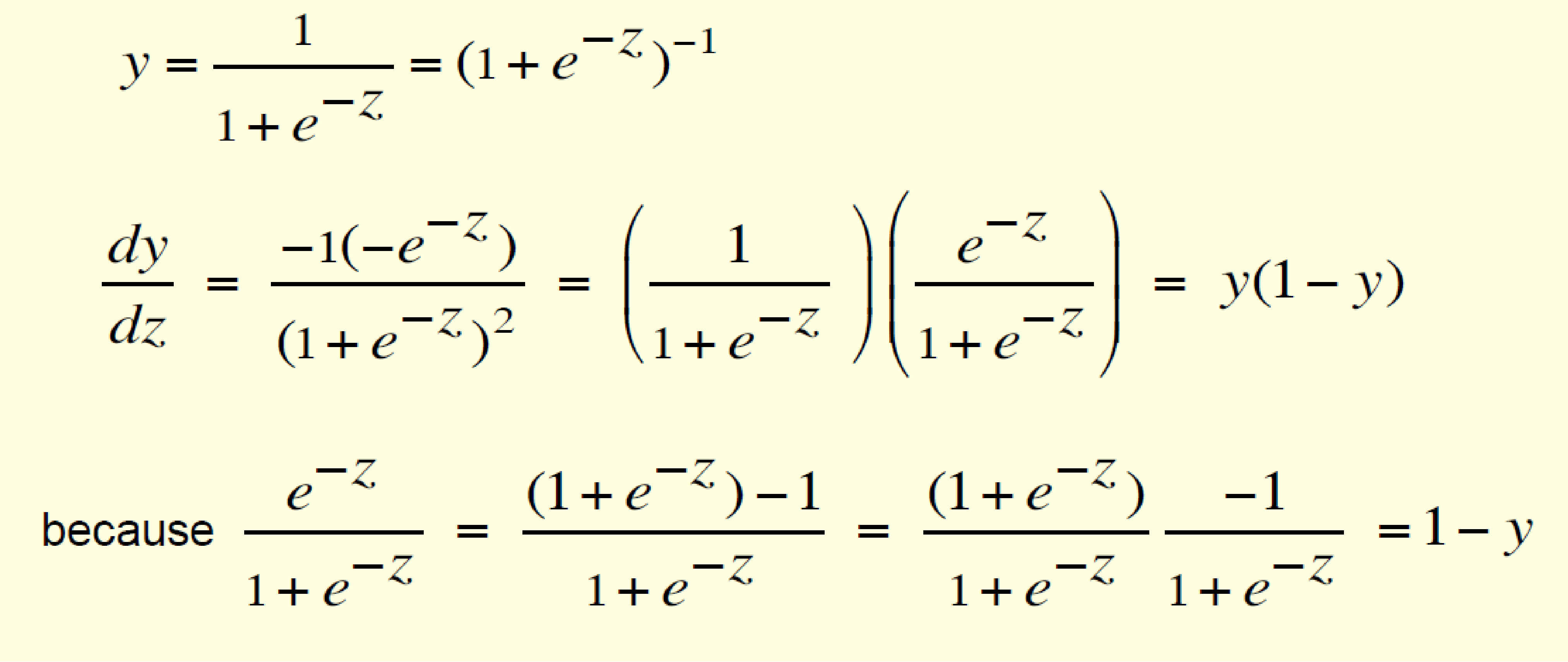


- These give a real-valued output that is a smooth and bounded function of their total input.
  - They have nice derivatives which make learning easy.

$$z = b + \sum_i x_i w_i \qquad y = \frac{1}{1+e^{-z}}$$



The derivatives of a logistic function

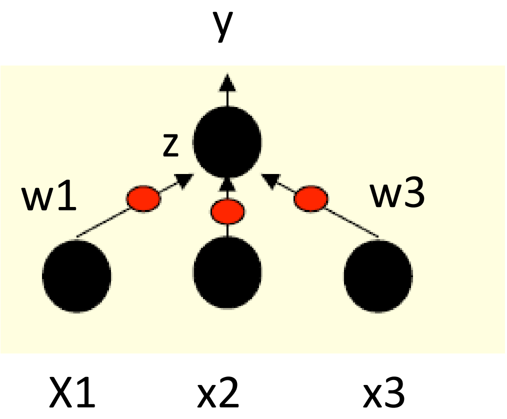


$$y = \frac{1}{1+e^{-z}} = (1+e^{-z})^{-1}$$

$$\frac{dy}{dz} = \frac{-1(-e^{-z})}{(1+e^{-z})^{2}} = \left(\frac{1}{1+e^{-z}}\right)\left(\frac{e^{-z}}{1+e^{-z}}\right) = y(1-y)$$

because $$\frac{e^{-z}}{1+e^{-z}} = \frac{(1+e^{-z})-1}{1+e^{-z}} = \frac{(1+e^{-z})}{1+e^{-z}}\,\frac{-1}{1+e^{-z}} = 1-y$$

Use the chain rule



- To learn the weights we need the derivative of the output with respect to each weight:

$$\frac{\partial y}{\partial w_i} = \frac{\partial z}{\partial w_i}\frac{dy}{dz} = x_i \; y \; (1-y)$$

delta-rule

$$\frac{\partial E}{\partial w_i} = \sum_n \frac{\partial y^n}{\partial w_i}\frac{\partial E}{\partial y^n} = -\sum_n \boxed{x_i^n}\;\boxed{y^n \; (1-y^n)}\;\boxed{(t^n - y^n)}$$

extra term = slope of logistic

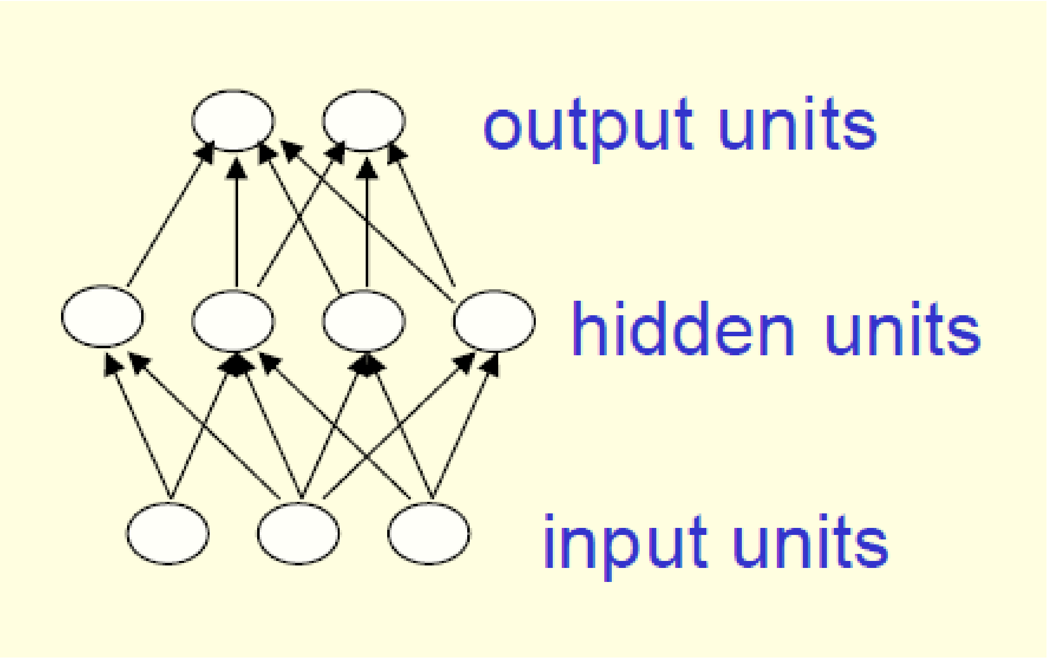
$$E = \sum_{n=1}^{N}(t^{(n)} - y^{(n)})^2$$

The superscript n here refers to n-th training instance.

Feed-forward neural network

- This is the simplest type of NN:
  - The first layer is the input and the last layer is the output
  - If there is more than one hidden layer, we call them deep NN

- Training: learn the weights on the arcs, using back propagation.



Learning with hidden units

- Networks without hidden units are very limited in the input-output mappings they can learn to model.

- More layers of linear units do not help. It’s still linear.

- We need multiple layers of adaptive, non-linear hidden units:
  - We need an efficient way of adapting all the weights, not just the last layer.
  - Learning the weights going into hidden units is equivalent to learning features.
  - This can be difficult because nobody is telling us directly what the hidden units should do.

Summary

- NN has input, output, and hidden layers.
- There are many kinds of neurons (i.e., using different activation functions).
- One important aspect of NNs is its non-linear activation functions.
- Learn the weights of NN using backpropagation.