

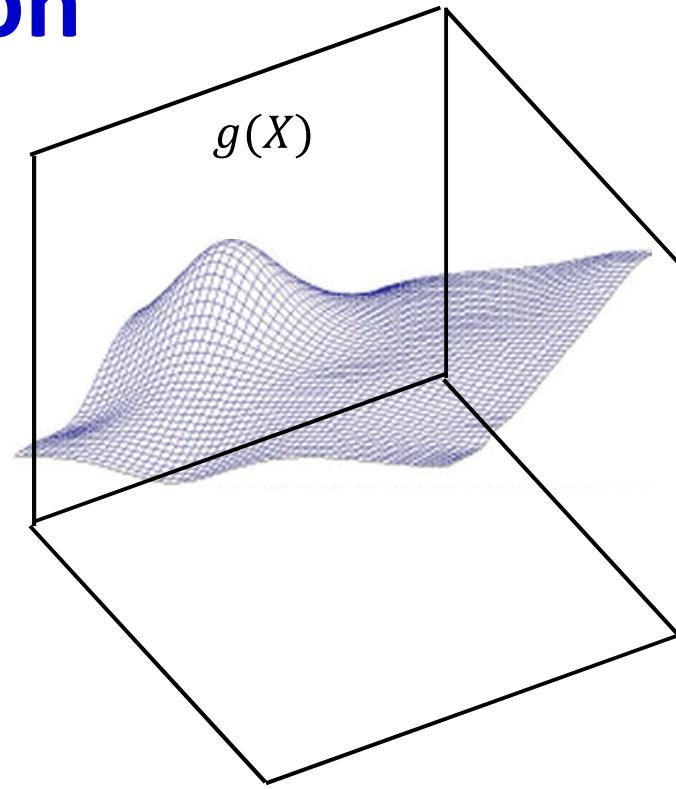
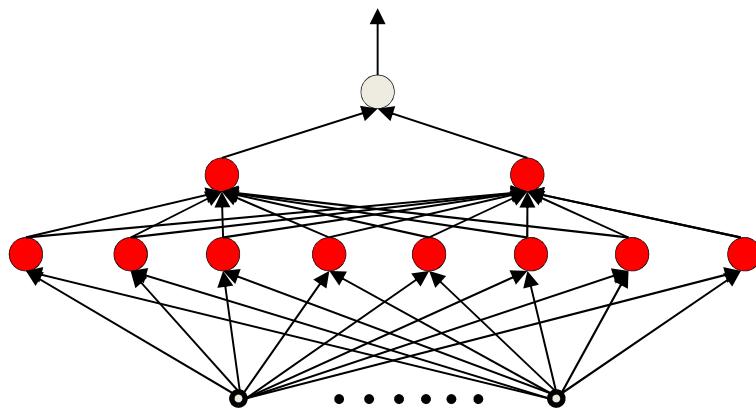
# **Neural Networks**

## **Learning the network: Backprop**

11-785, Spring 2020

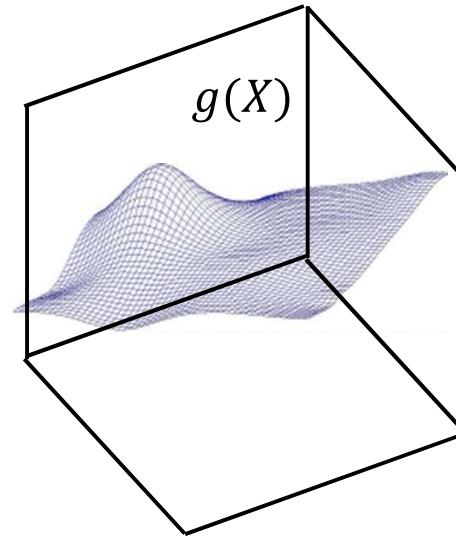
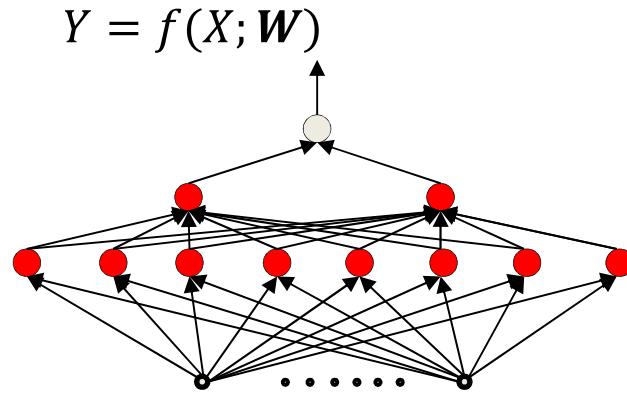
Lecture 4

# Recap: The MLP *can* represent any function



- The MLP *can be constructed* to represent anything
- But *how* do we construct it?
  - *I.e.* how do we determine the weights (and biases) of the network to best represent a target function
    - *Assuming that the architecture of the network is given*

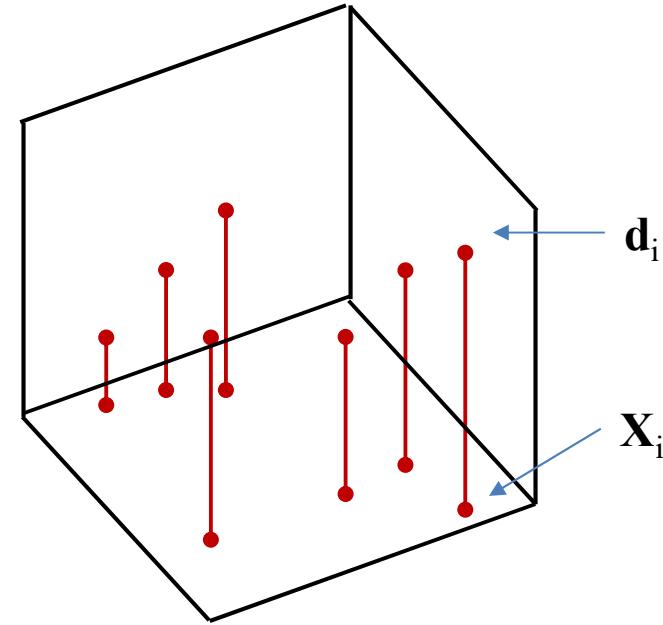
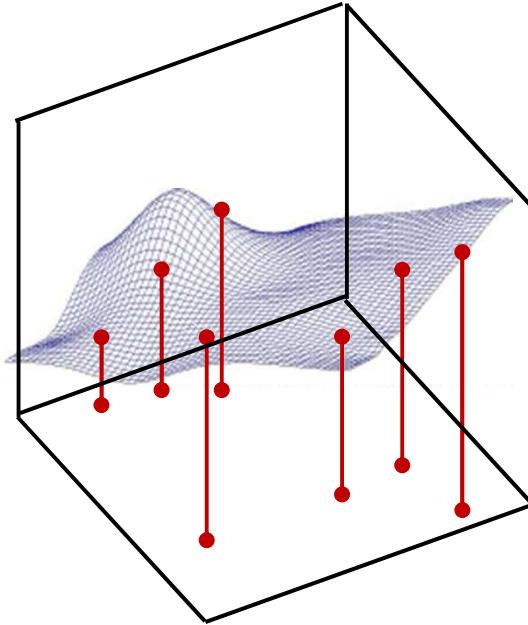
# Recap: How to learn the function



- By minimizing expected error

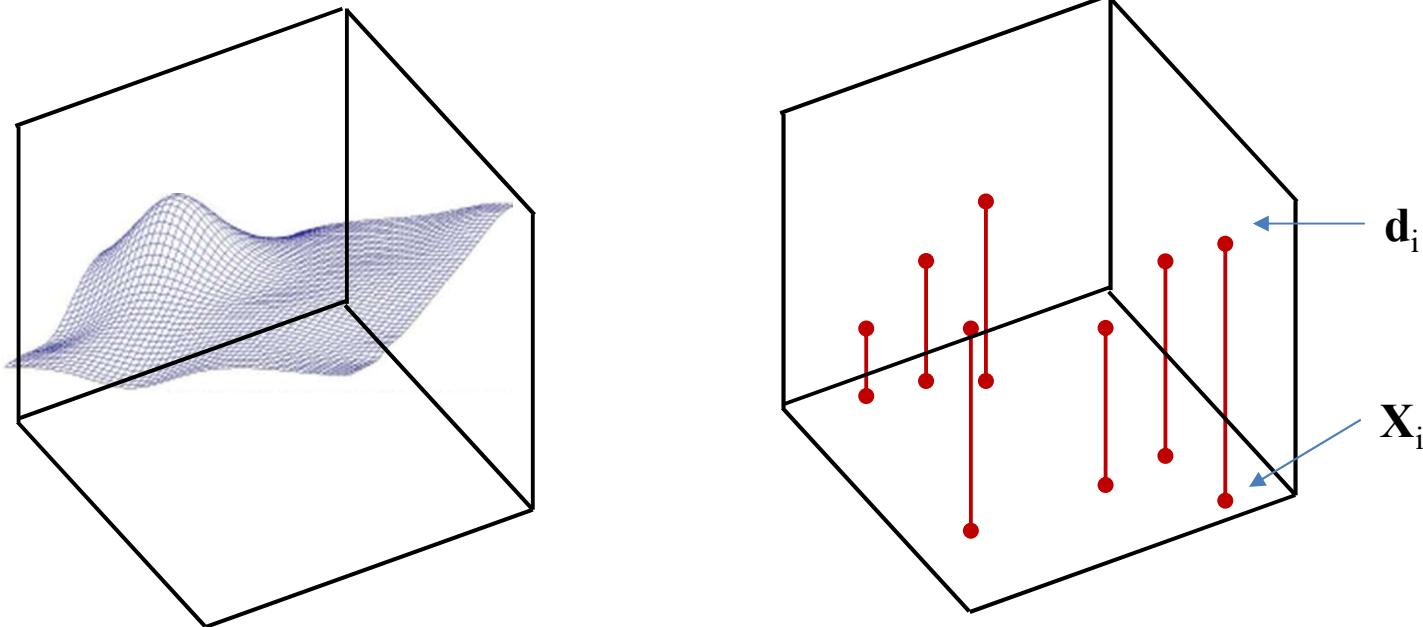
$$\begin{aligned}\widehat{W} &= \operatorname{argmin}_W \int_X \operatorname{div}(f(X; W), g(X)) P(X) dX \\ &= \operatorname{argmin}_W E[\operatorname{div}(f(X; W), g(X))]\end{aligned}$$

# Recap: Sampling the function



- $g(X)$  is unknown, so sample it
  - Basically, get input-output pairs for a number of samples of input  $X_i$
  - Good sampling: the samples of  $X$  will be drawn from  $P(X)$
- Estimate function from the samples

# The *Empirical* risk

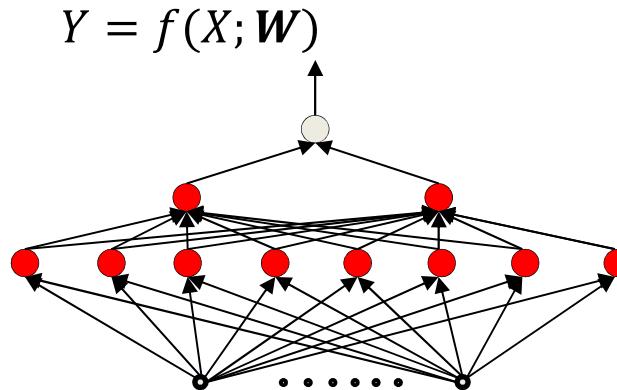


- The *empirical estimate* of the expected error is the *average* error over the samples

$$E[\text{div}(f(X; W), g(X))] \approx \frac{1}{T} \sum_{i=1}^T \text{div}(f(X_i; W), d_i)$$

- This approximation is an unbiased estimate of the *expected* divergence that we *actually* want to estimate
  - We can *hope* that minimizing the empirical loss will minimize the true loss
  - Caveat: This hope is generally not based on anything but, well, hope..

# Empirical Risk Minimization



- Given a training set of input-output pairs  $(\mathbf{X}_1, \mathbf{d}_1), (\mathbf{X}_2, \mathbf{d}_2), \dots, (\mathbf{X}_T, \mathbf{d}_T)$ 
  - Error on the  $i$ -th instance:  $\text{div}(f(\mathbf{X}_i; \mathbf{W}), d_i)$
  - Empirical average error on all training data:

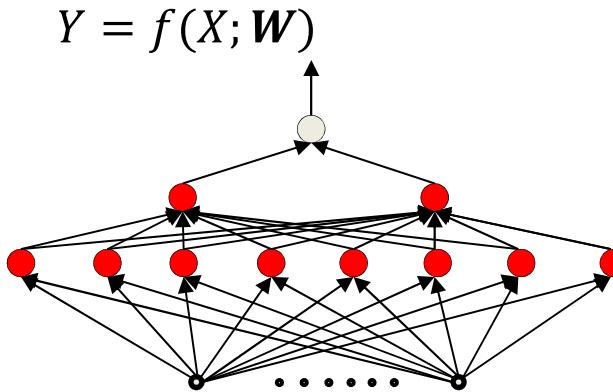
$$\text{Loss}(\mathbf{W}) = \frac{1}{T} \sum_i \text{div}(f(\mathbf{X}_i; \mathbf{W}), d_i)$$

- Estimate the parameters to minimize the empirical estimate of expected error

$$\widehat{\mathbf{W}} = \operatorname{argmin}_{\mathbf{W}} \text{Loss}(\mathbf{W})$$

- I.e. minimize the *empirical error* over the drawn samples

# Empirical Risk Minimization



This is an instance of  
function minimization  
(optimization)

- Given a training set of input-output pairs  $(\mathbf{X}_1, \mathbf{d}_1), (\mathbf{X}_2, \mathbf{d}_2), \dots, (\mathbf{X}_T, \mathbf{d}_T)$ 
  - Error on the  $i$ -th instance:  $\text{div}(f(\mathbf{X}_i; \mathbf{W}), \mathbf{d}_i)$
  - Empirical average error on all training data:

$$\text{Loss}(\mathbf{W}) = \frac{1}{T} \sum_i \text{div}(f(\mathbf{X}_i; \mathbf{W}), \mathbf{d}_i)$$

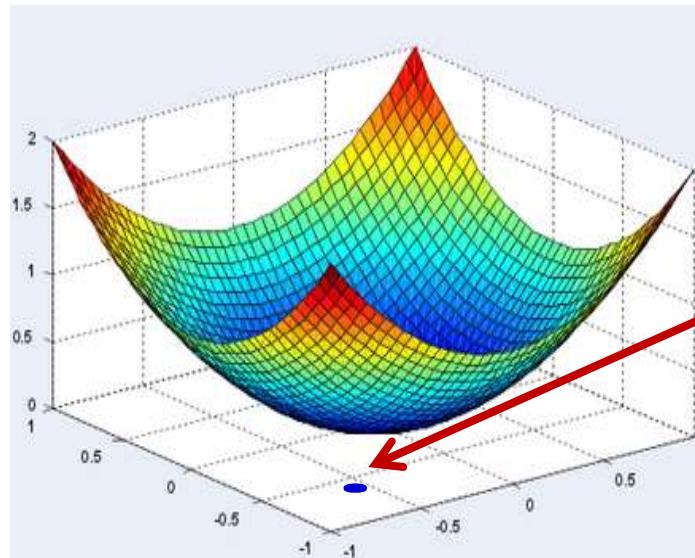
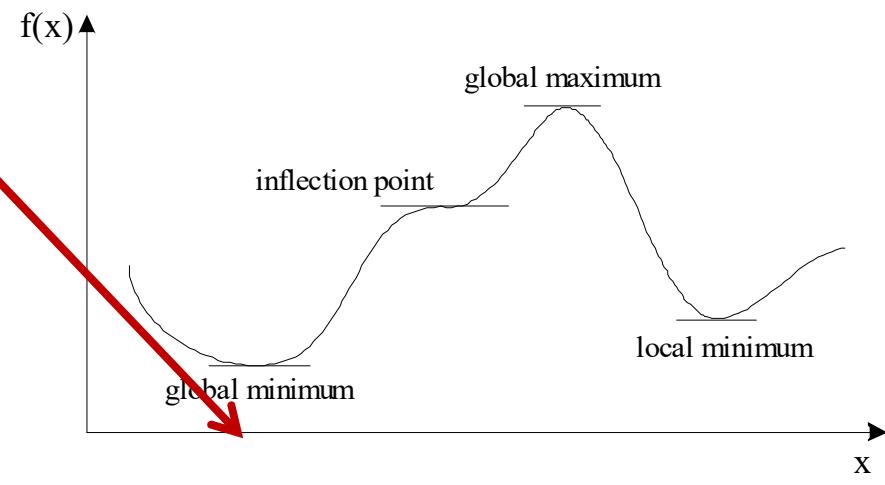
- Estimate the parameters to minimize the empirical estimate of expected error

$$\widehat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \text{Loss}(\mathbf{W})$$

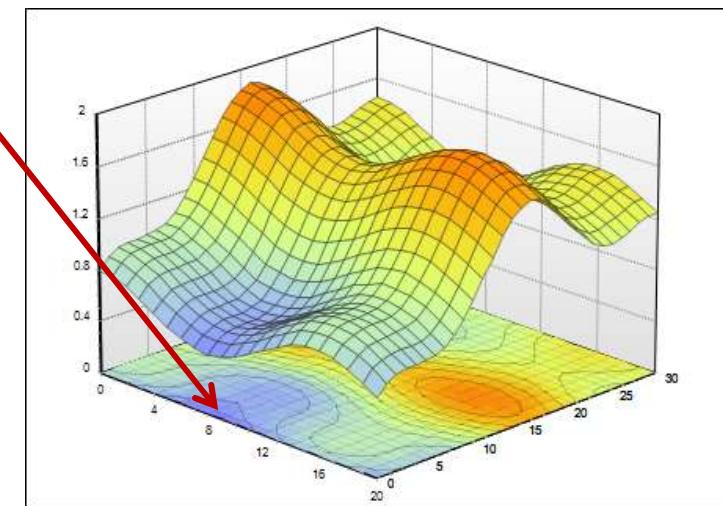
- I.e. minimize the *empirical error* over the drawn samples

- A CRASH COURSE ON FUNCTION  
OPTIMIZATION

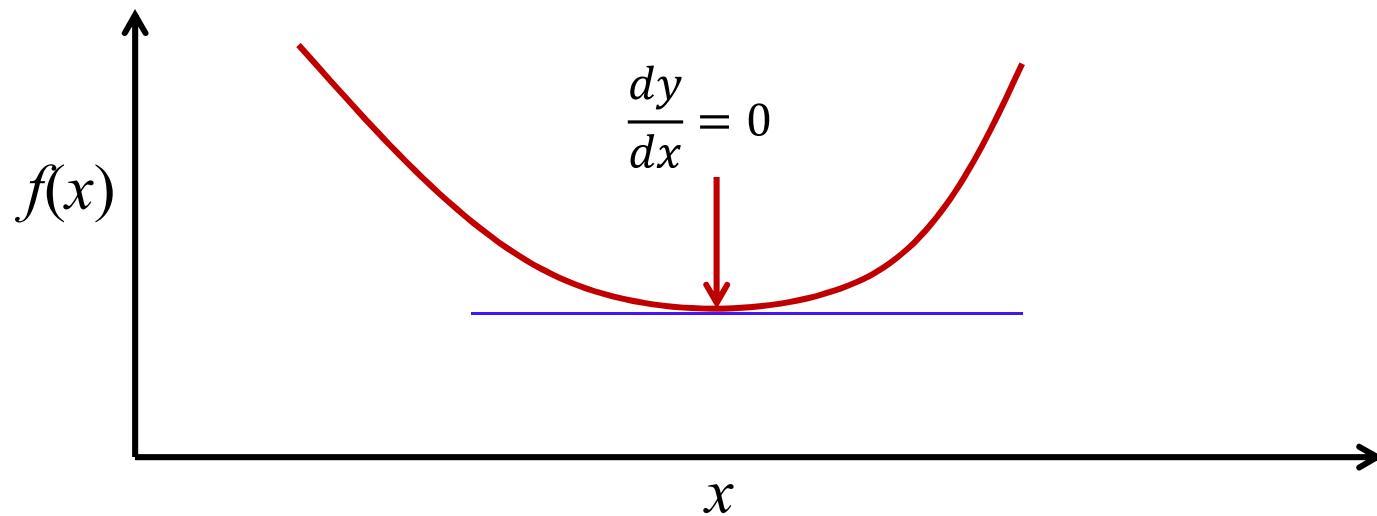
# The problem of optimization



- General problem of optimization: find the value of  $x$  where  $f(x)$  is minimum

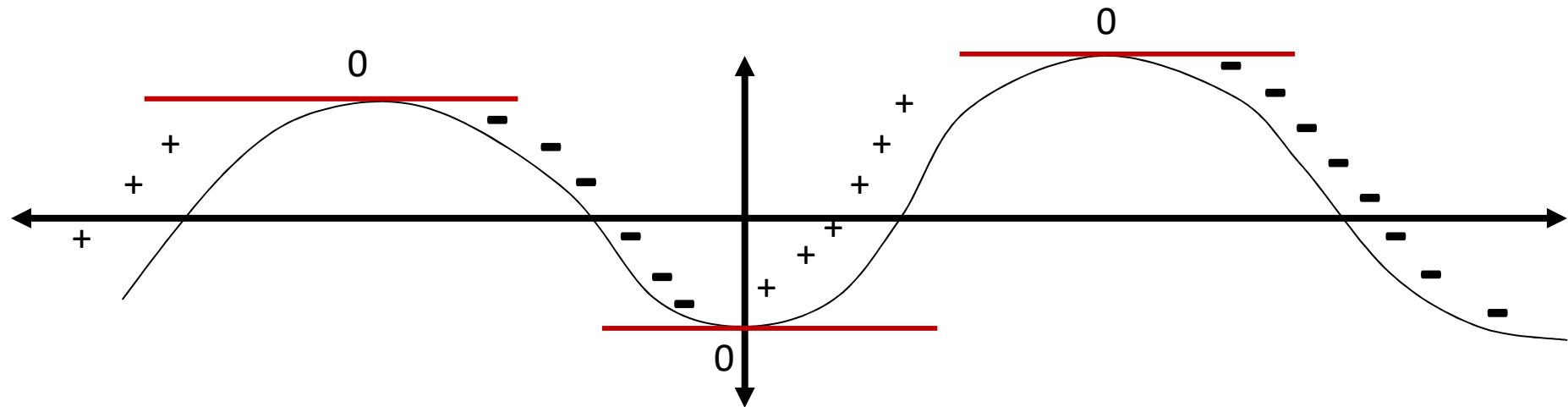


# Finding the minimum of a function



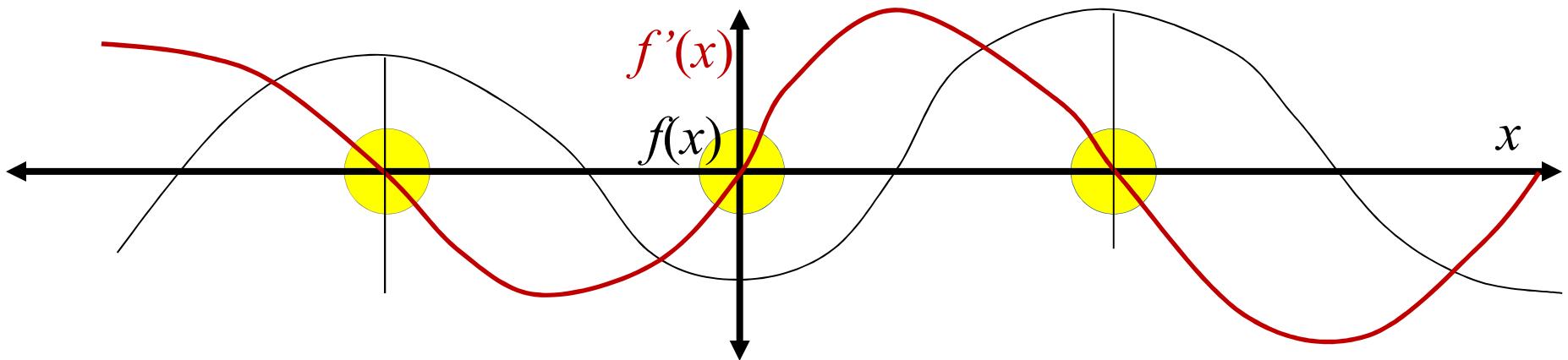
- Find the value  $x$  at which  $f'(x) = 0$ 
  - Solve
- The solution is a “turning point”
  - Derivatives go from positive to negative or vice versa at this point
- But is it a minimum?

# Turning Points



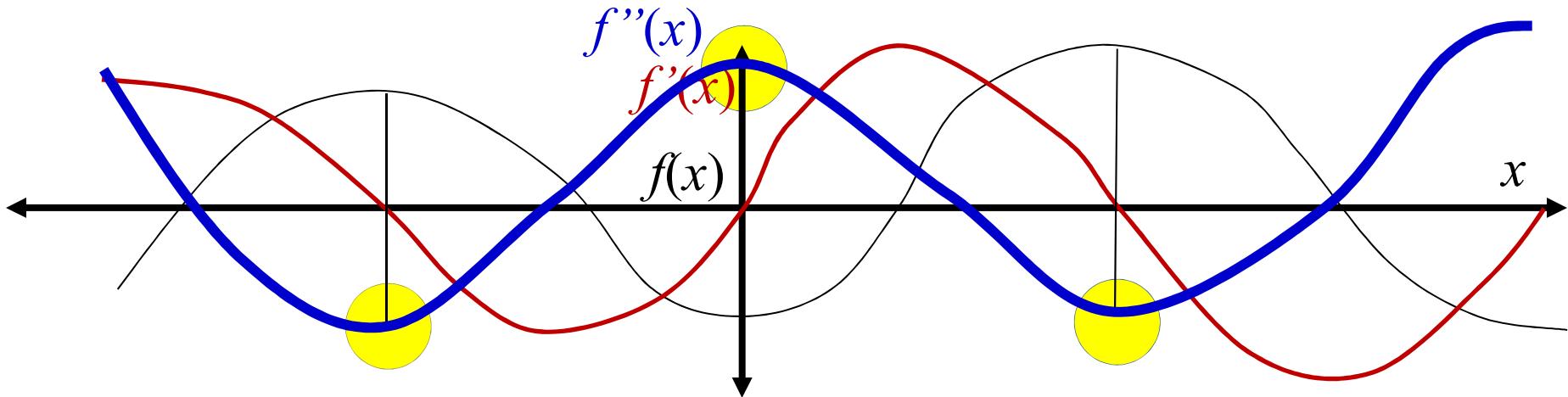
- Both *maxima* and *minima* have zero derivative
- Both are turning points

# Derivatives of a curve



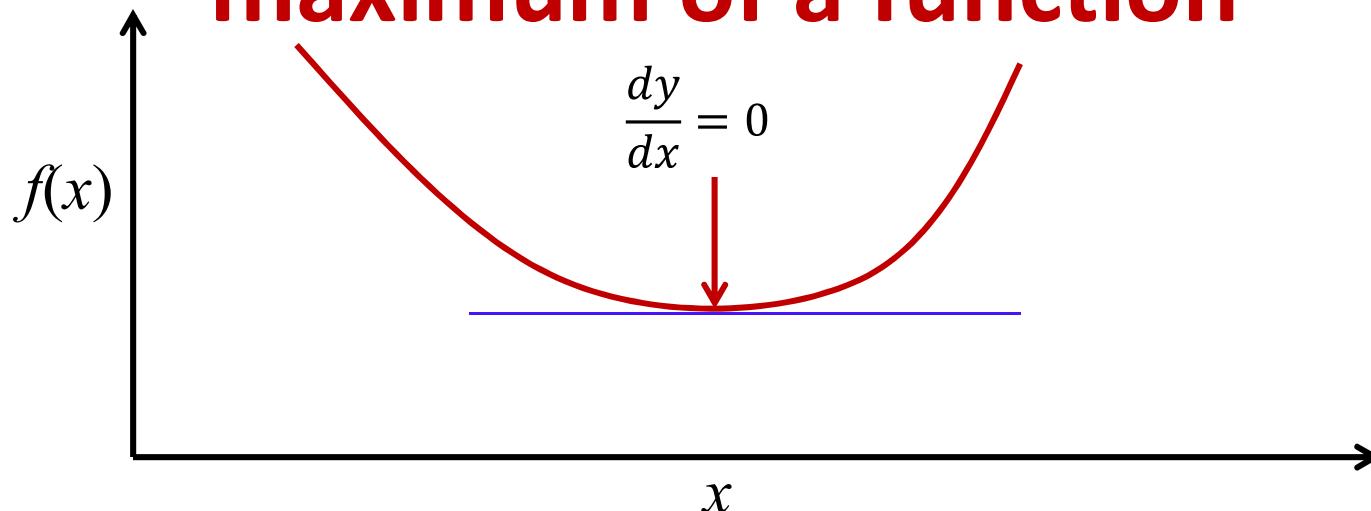
- Both *maxima* and *minima* are turning points
- Both *maxima* and *minima* have **zero derivative**

# Derivative of the derivative of the curve



- Both *maxima* and *minima* are turning points
- Both *maxima* and *minima* have zero derivative
- The *second derivative*  $f''(x)$  is –ve at maxima and +ve at minima!

# Soln: Finding the minimum or maximum of a function



- Find the value  $x$  at which  $f'(x) = 0$ : Solve

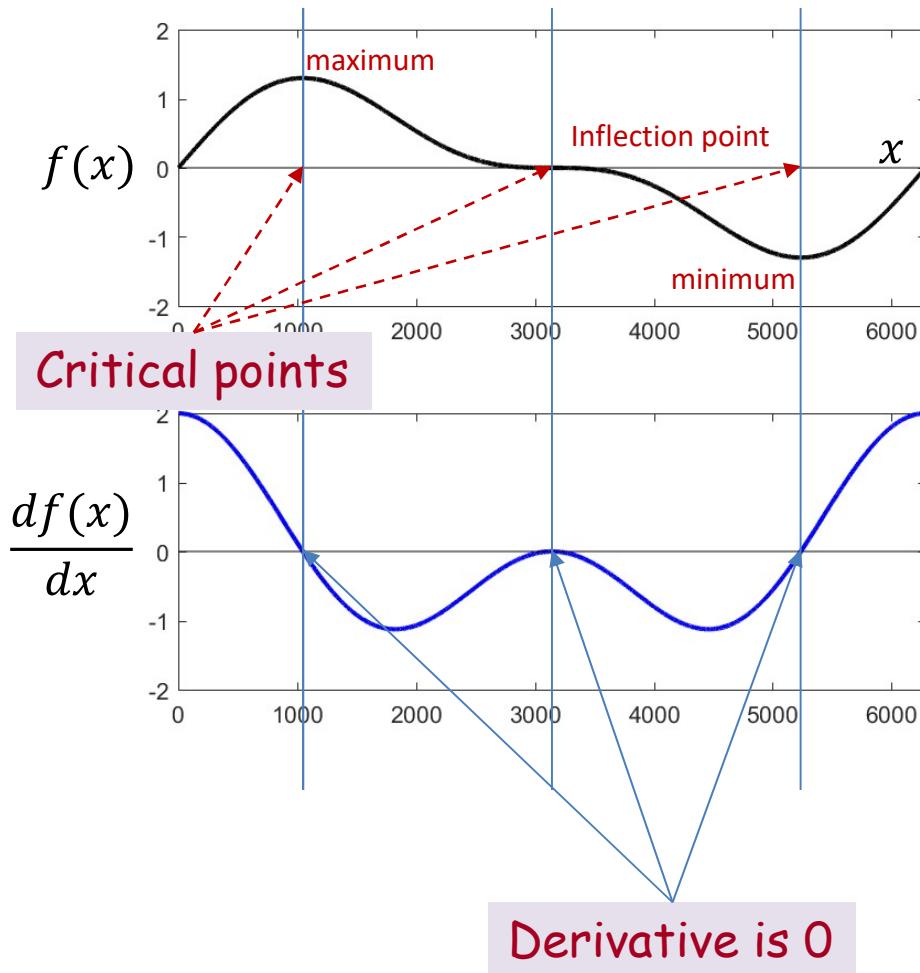
$$\frac{df(x)}{dx} = 0$$

- The solution  $x_{soln}$  is a turning point
- Check the double derivative at  $x_{soln}$  : compute

$$f''(x_{soln}) = \frac{d^2f(x_{soln})}{dx^2}$$

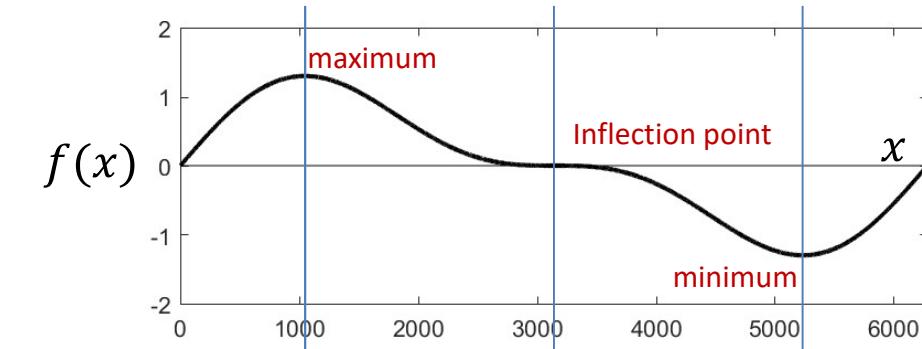
- If  $f''(x_{soln})$  is positive  $x_{soln}$  is a minimum, otherwise it is a maximum

# A note on derivatives of functions of single variable

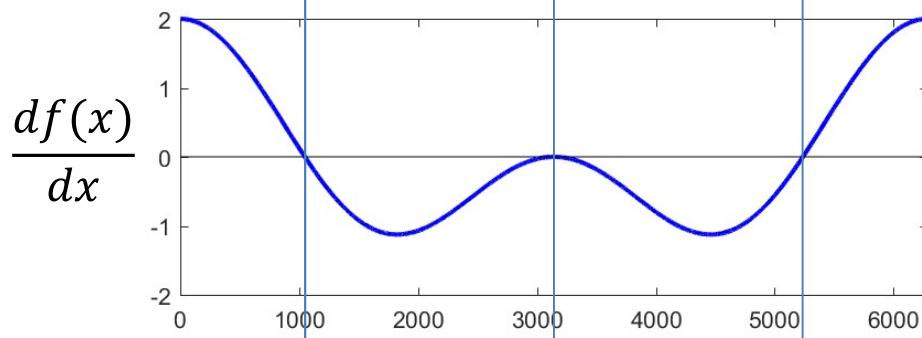


- All locations with zero derivative are *critical* points
  - These can be local maxima, local minima, or inflection points

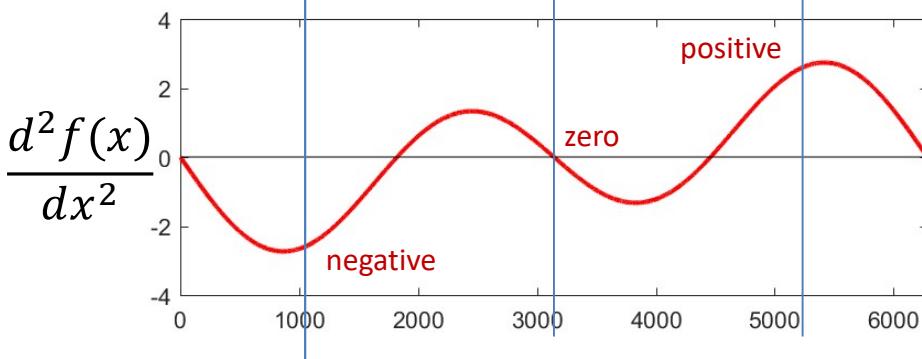
# A note on derivatives of functions of single variable



- All locations with zero derivative are *critical* points
  - These can be local maxima, local minima, or inflection points

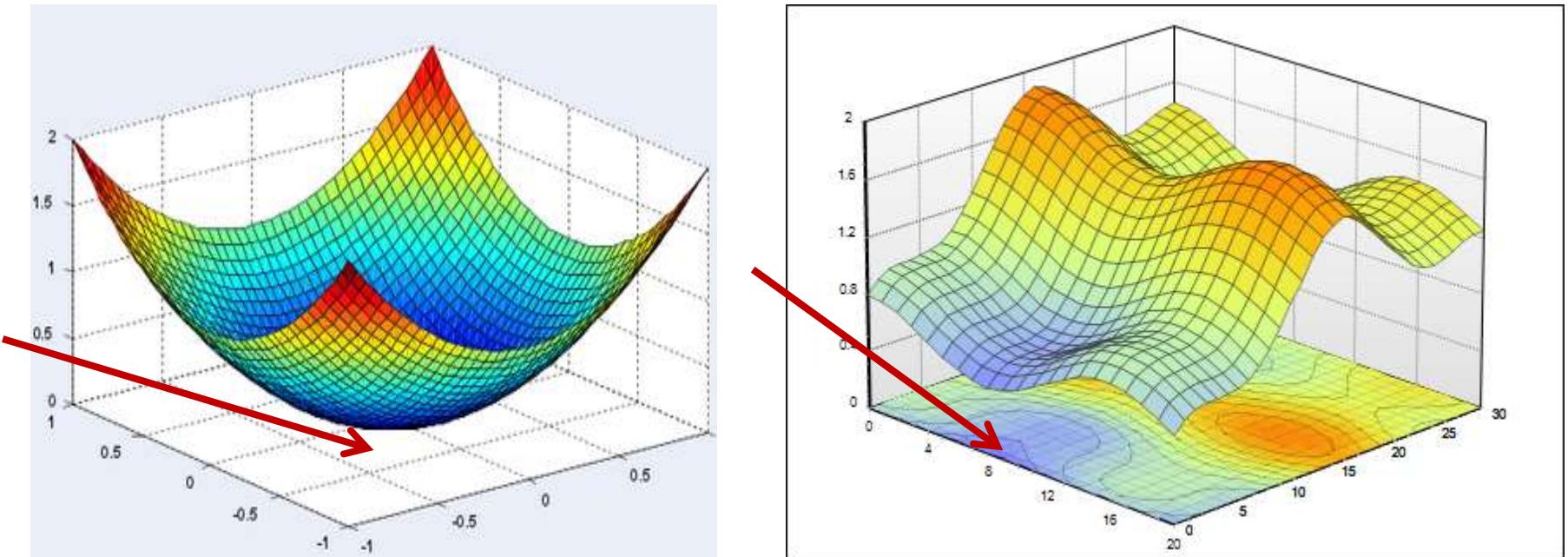


- The *second* derivative is
  - $\geq 0$  at minima
  - $\leq 0$  at maxima
  - Zero at inflection points



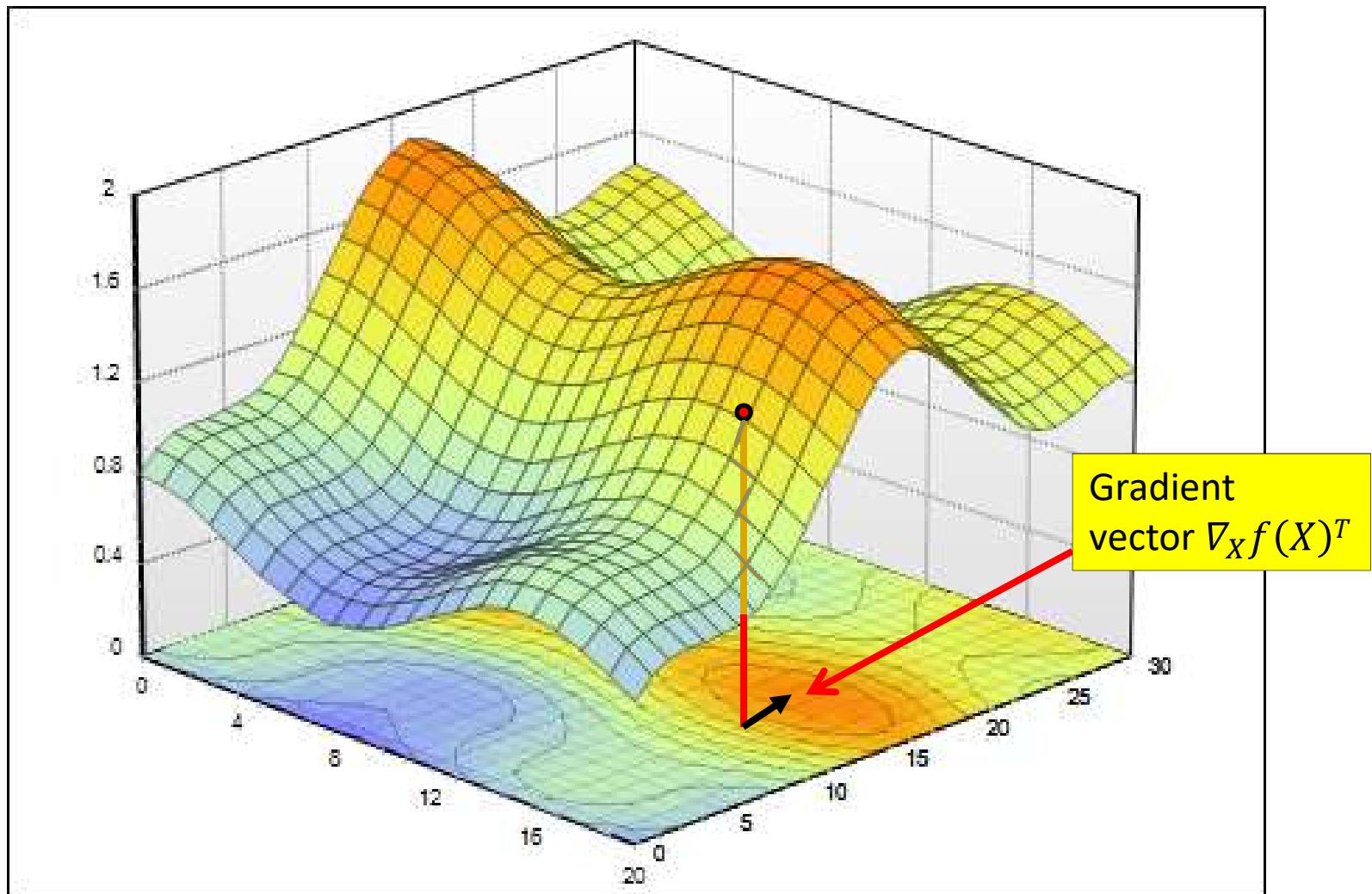
- It's a little more complicated for functions of multiple variables..

# What about functions of multiple variables?



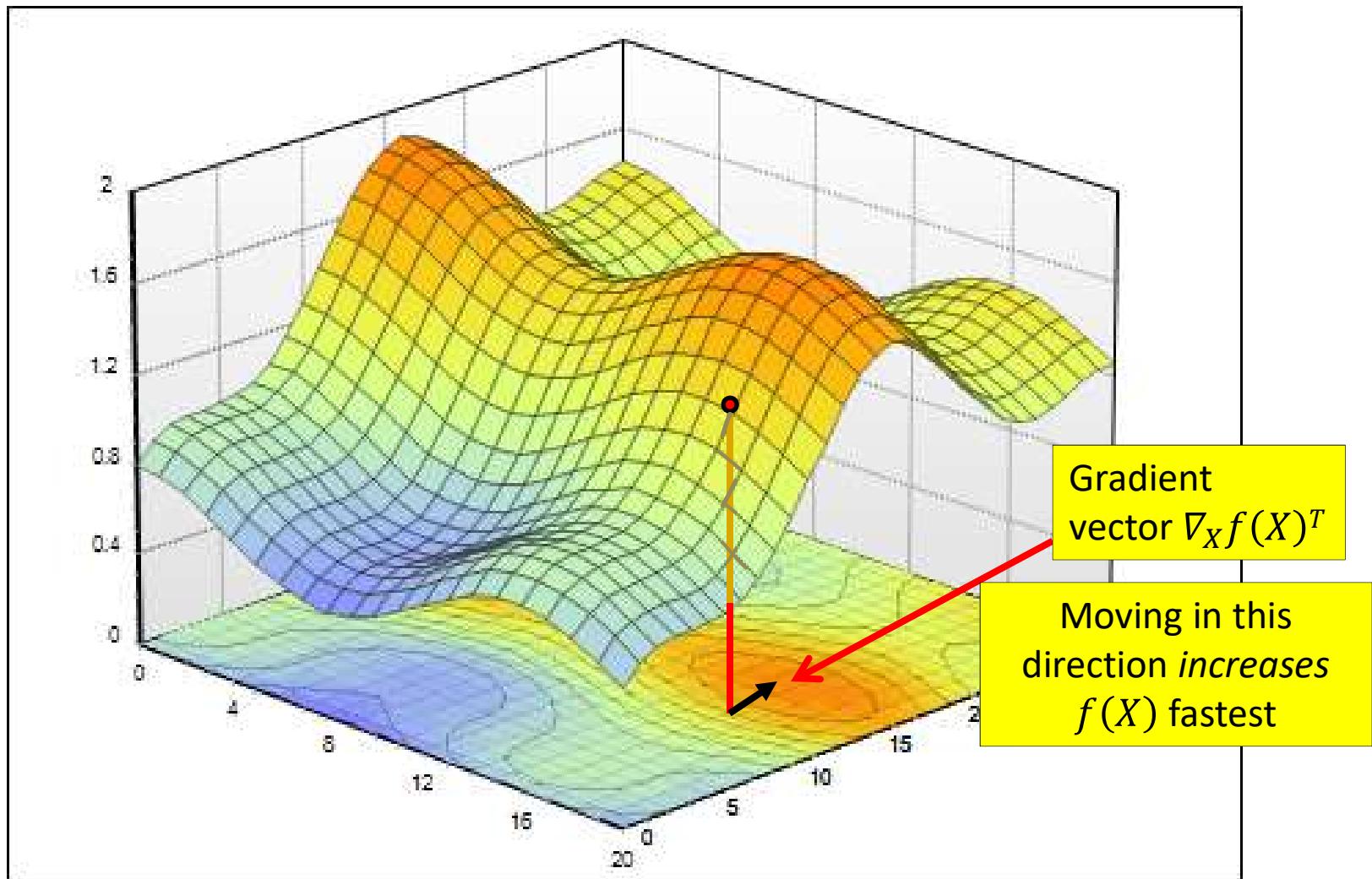
- The optimum point is still “turning” point
  - Shifting in any direction will increase the value
  - For smooth functions, minuscule shifts will not result in any change at all
- We must find a point where shifting in any direction by a microscopic amount will not change the value of the function

# Gradient

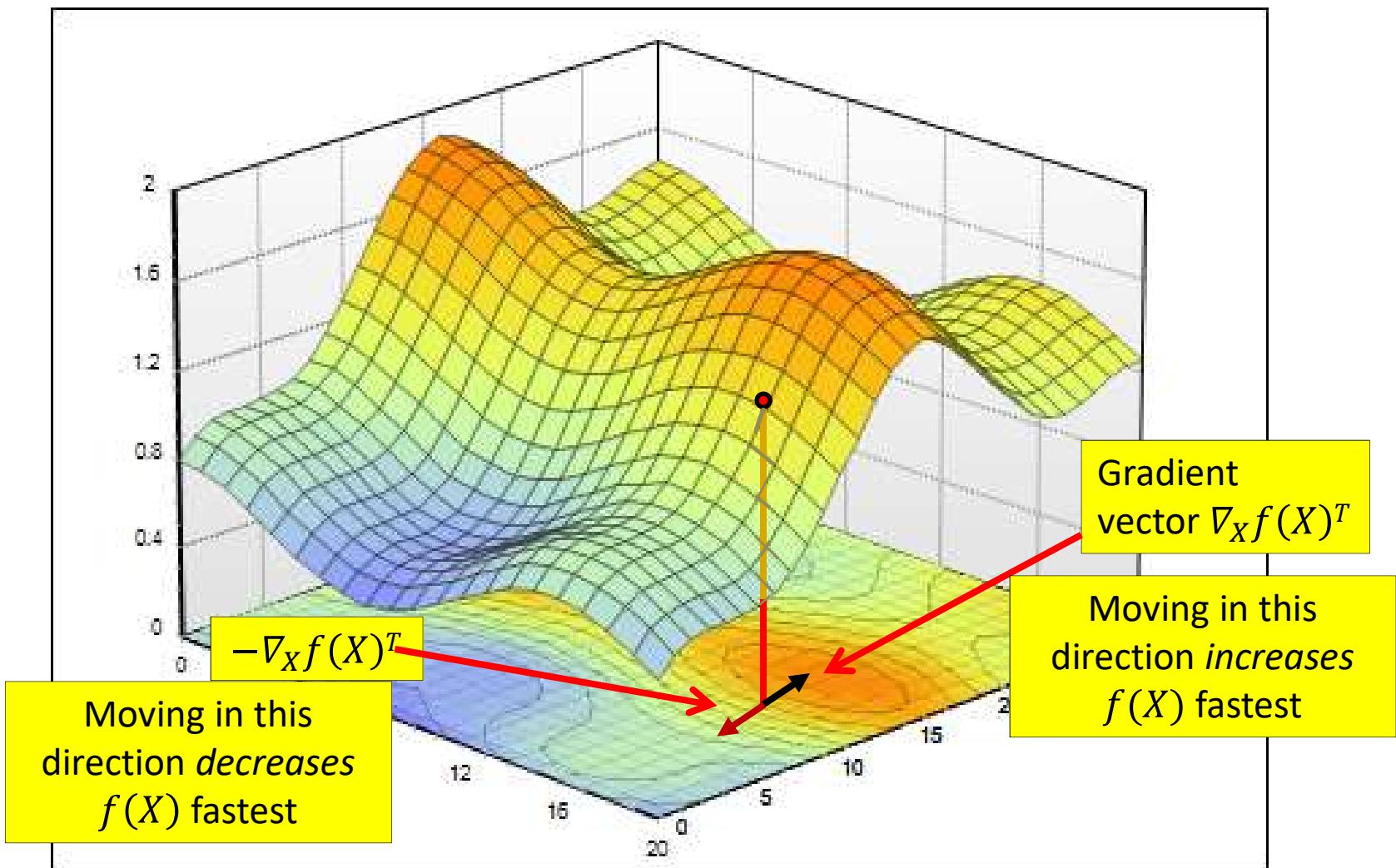


The gradient is the direction of fastest increase of the function

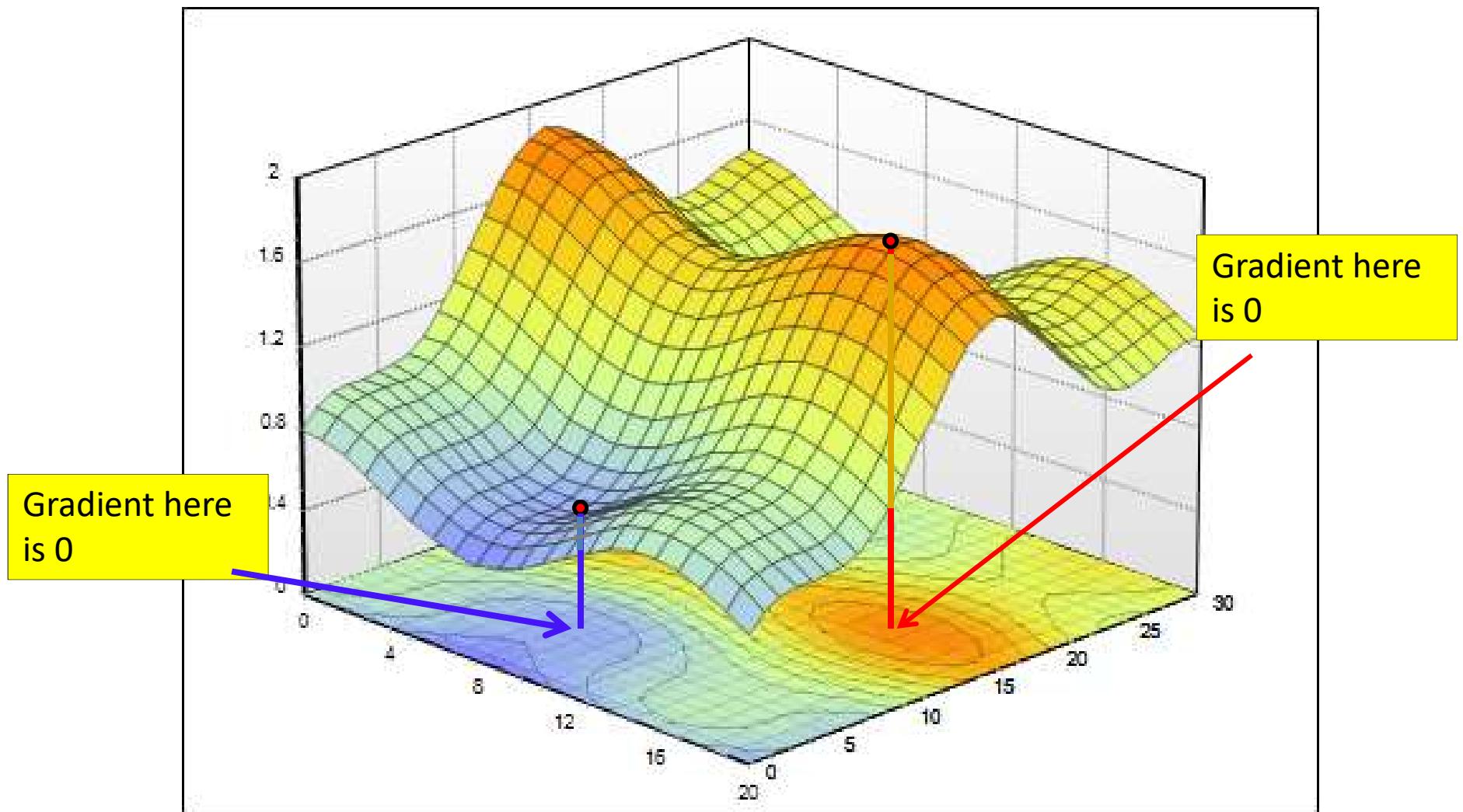
# Gradient



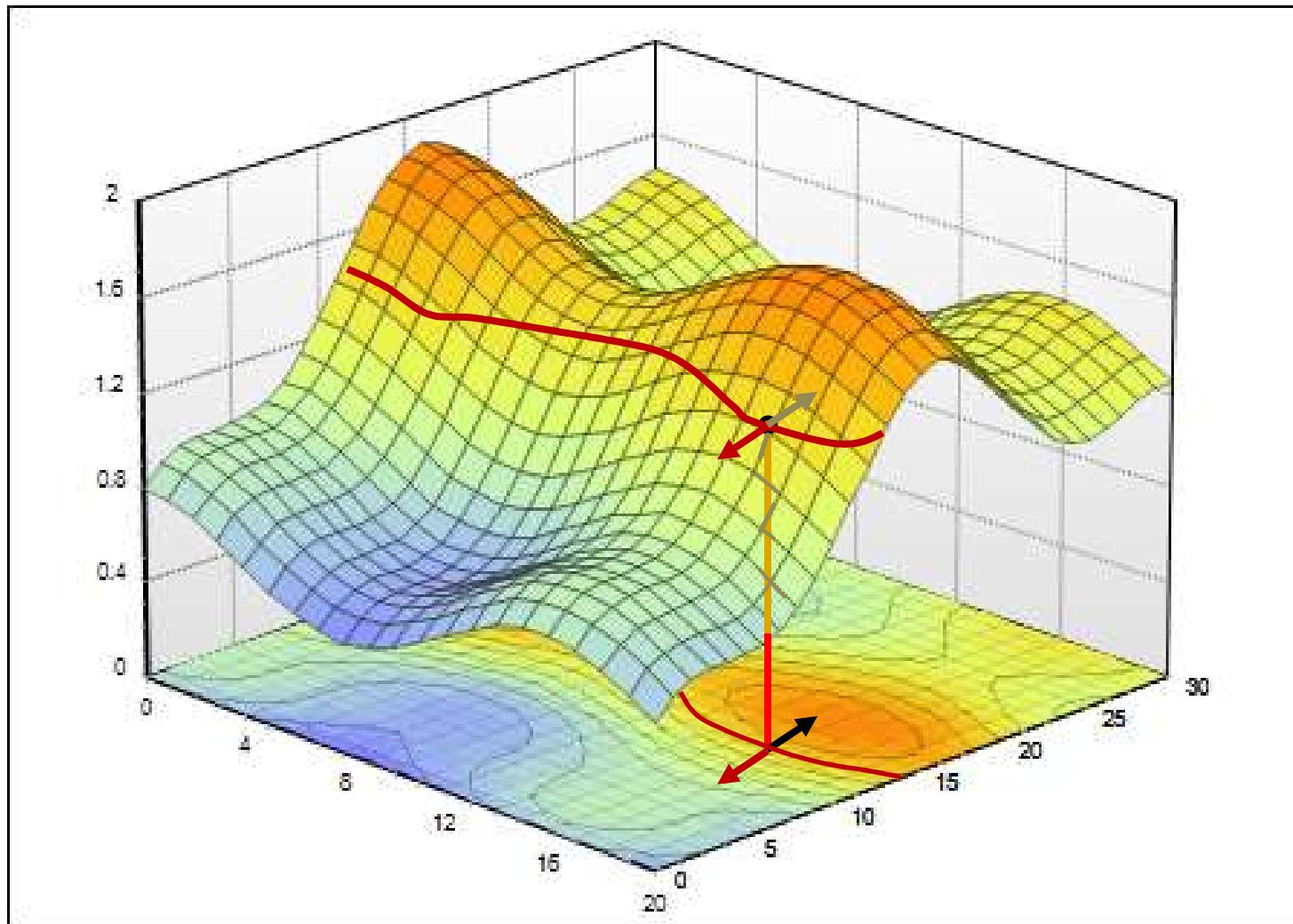
# Gradient



# Gradient



# Properties of Gradient: 2



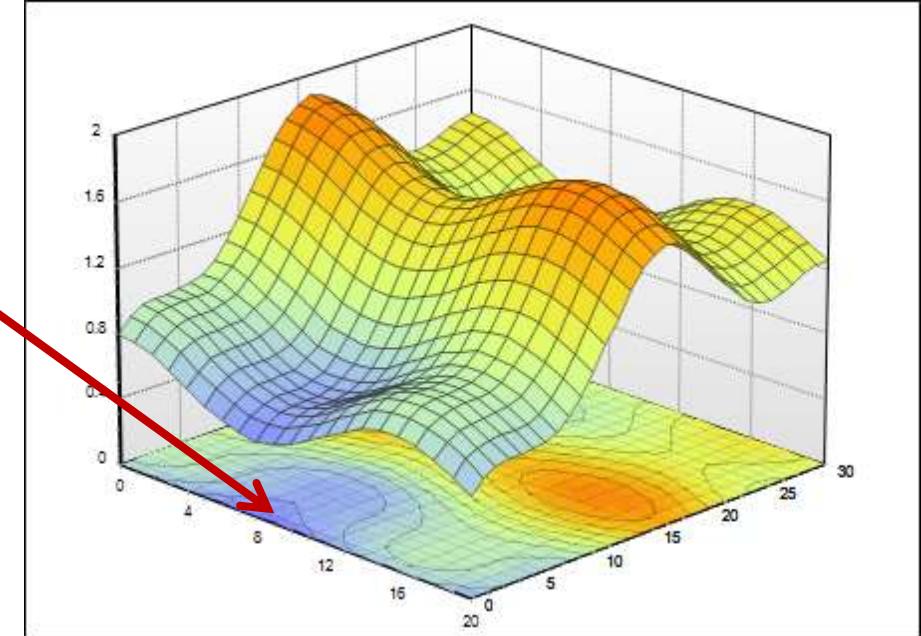
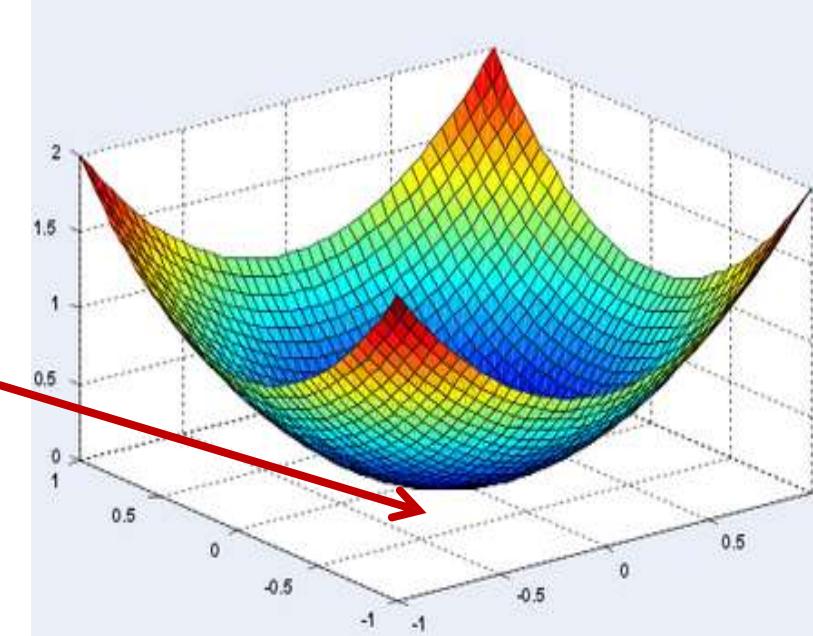
- The gradient vector  $\nabla_X f(X)^T$  is perpendicular to the level curve

# The Hessian

- The Hessian of a function  $f(x_1, x_2, \dots, x_n)$  is given by the second derivative

$$\nabla_x^2 f(x_1, \dots, x_n) := \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdot & \cdot & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdot & \cdot & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdot & \cdot & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

# Finding the minimum of a scalar function of a multi-variate input



- The optimum point is a turning point – the gradient will be 0

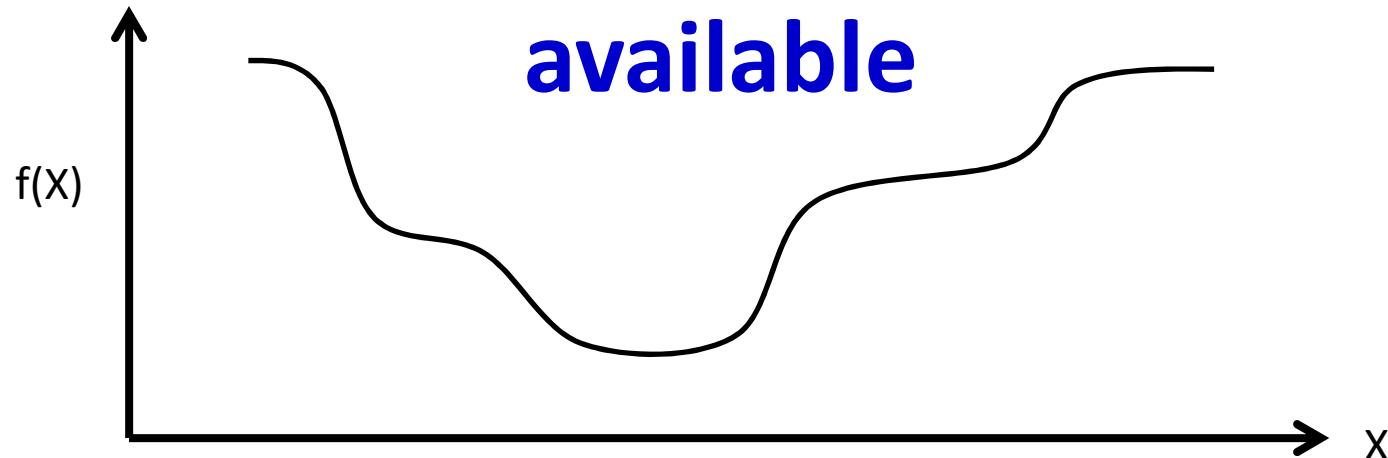
# Unconstrained Minimization of function (Multivariate)

1. Solve for the  $X$  where the derivative (or gradient) equals to zero

$$\nabla_X f(X) = 0$$

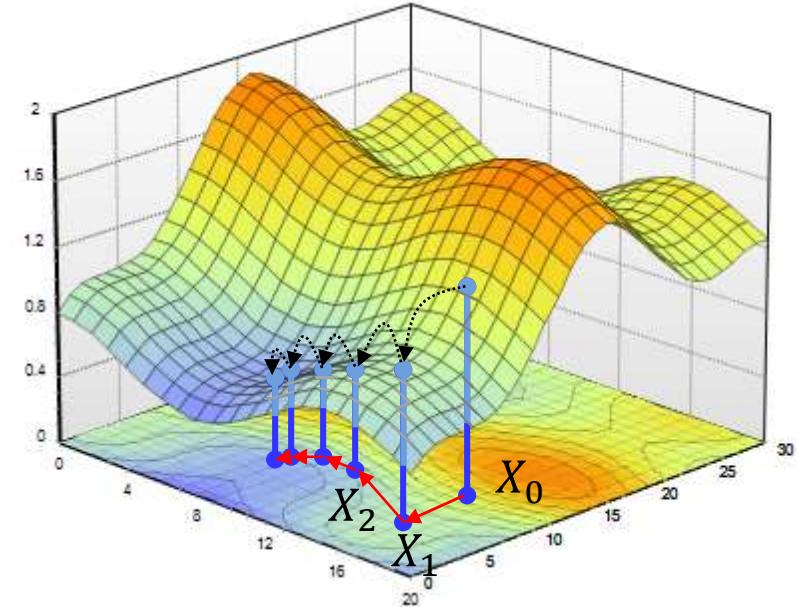
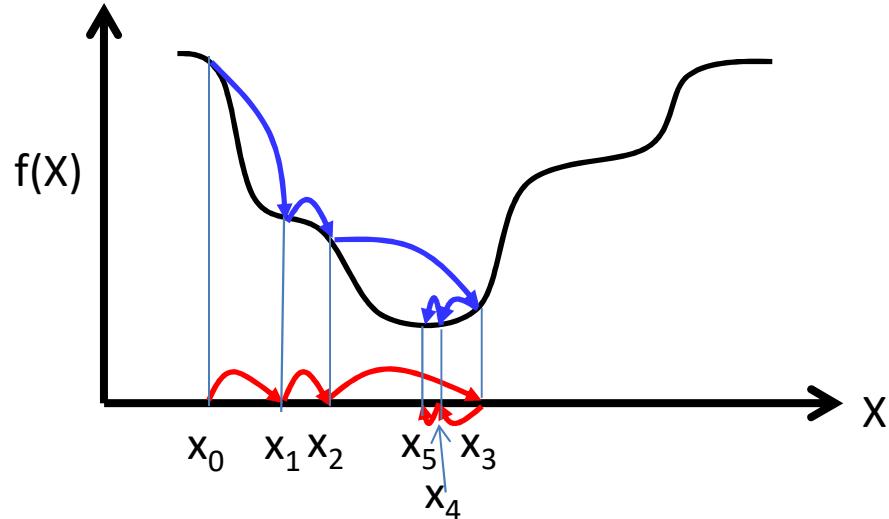
2. Compute the Hessian Matrix  $\nabla_X^2 f(X)$  at the candidate solution and verify that
  - Hessian is positive definite (eigenvalues positive) -> to identify local minima
  - Hessian is negative definite (eigenvalues negative) -> to identify local maxima

# Closed Form Solutions are not always available



- Often it is not possible to simply solve  $\nabla_X f(X) = 0$ 
  - The function to minimize/maximize may have an intractable form
- In these situations, iterative solutions are used
  - Begin with a “guess” for the optimal  $X$  and refine it iteratively until the correct value is obtained

# Iterative solutions



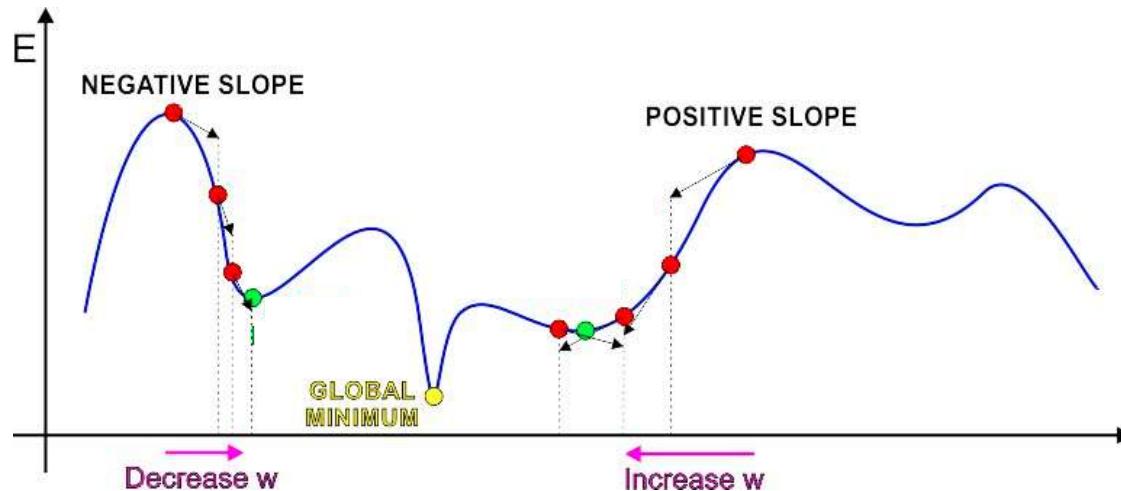
- Iterative solutions
  - Start from an initial guess  $X_0$  for the optimal  $X$
  - Update the guess towards a (hopefully) “better” value of  $f(X)$
  - Stop when  $f(X)$  no longer decreases
- Problems:
  - Which direction to step in
  - How big must the steps be

# The Approach of Gradient Descent



- Iterative solution:
  - Start at some point
  - Find direction in which to shift this point to decrease error
    - This can be found from the derivative of the function
      - A positive derivative → moving left decreases error
      - A negative derivative → moving right decreases error
  - Shift point in this direction

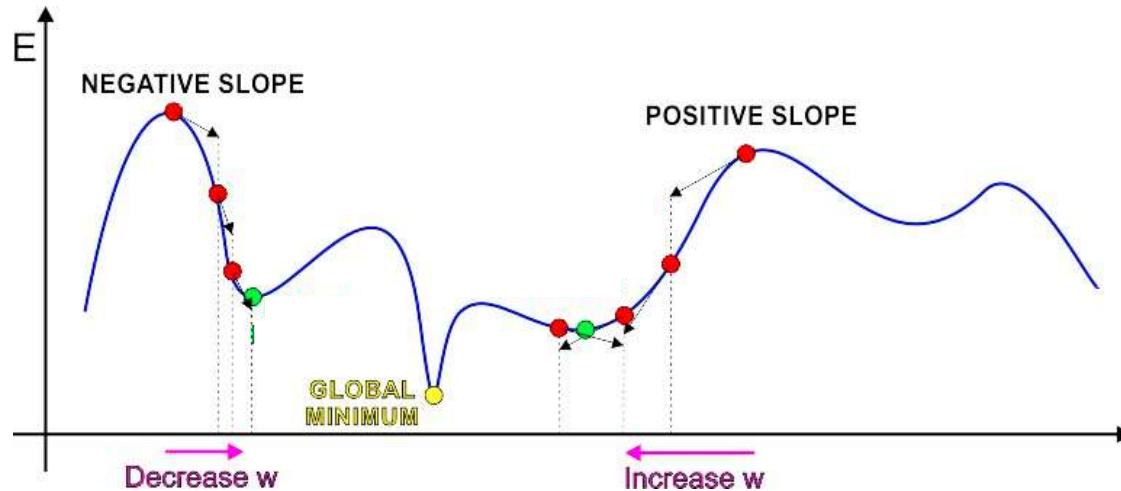
# The Approach of Gradient Descent



- Iterative solution: Trivial algorithm

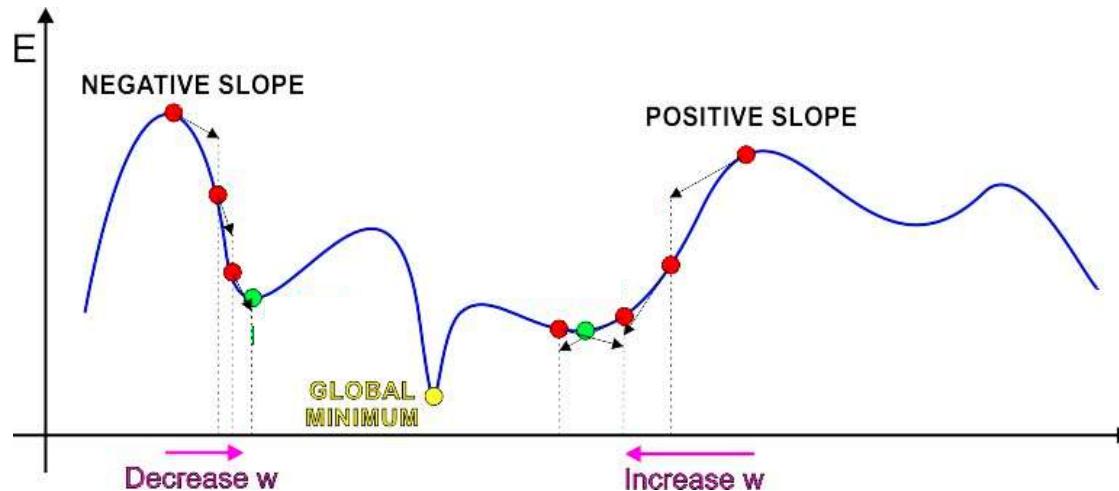
- Initialize  $x^0$
- While  $f'(x^k) \neq 0$ 
  - If  $\text{sign}(f'(x^k))$  is positive:  
$$x^{k+1} = x^k - \text{step}$$
  - Else  
$$x^{k+1} = x^k + \text{step}$$

# The Approach of Gradient Descent



- Iterative solution: Trivial algorithm
  - Initialize  $x^0$
  - While  $f'(x^k) \neq 0$ 
$$x^{k+1} = x^k - \text{sign}(f'(x^k)).step$$
- Identical to previous algorithm

# The Approach of Gradient Descent



- Iterative solution: Trivial algorithm
  - Initialize  $x^0$
  - While  $f'(x^k) \neq 0$ 
$$x^{k+1} = x^k - \eta^k f'(x^k)$$
- $\eta^k$  is the “step size”

# Gradient descent/ascent (multivariate)

- The gradient descent/ascent method to find the minimum or maximum of a function  $f$  iteratively

- To find a *maximum* move *in the direction of the gradient*

$$x^{k+1} = x^k + \eta^k \nabla_x f(x^k)^T$$

- To find a *minimum* move *exactly opposite the direction of the gradient*

$$x^{k+1} = x^k - \eta^k \nabla_x f(x^k)^T$$

- Many solutions to choosing step size  $\eta^k$

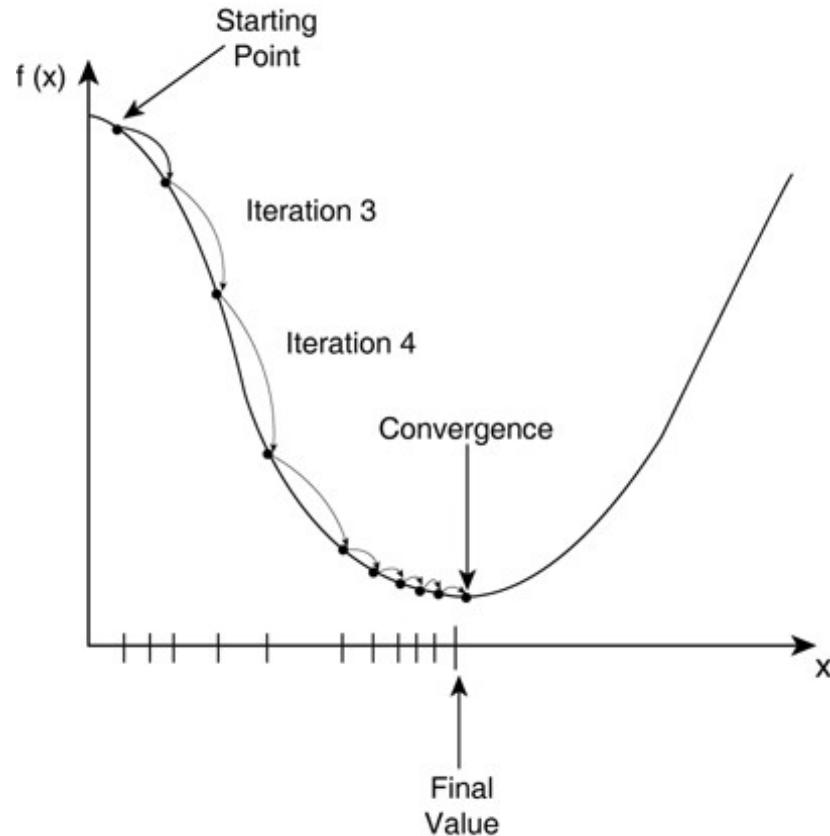
# Gradient descent convergence criteria

- The gradient descent algorithm converges when one of the following criteria is satisfied

$$|f(x^{k+1}) - f(x^k)| < \varepsilon_1$$

- Or

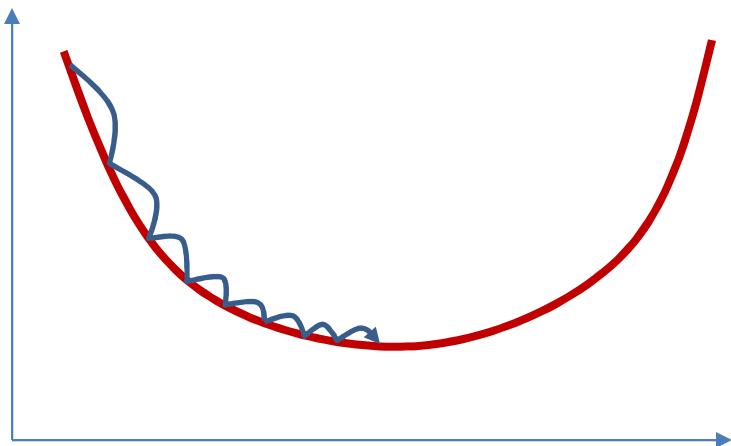
$$\|\nabla_x f(x^k)\| < \varepsilon_2$$



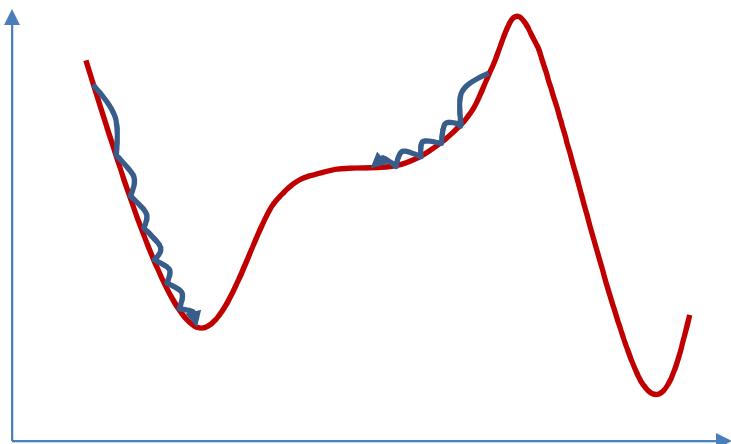
# Overall Gradient Descent Algorithm

- Initialize:
  - $x^0$
  - $k = 0$
- do
  - $x^{k+1} = x^k - \eta^k \nabla_x f(x^k)^T$
  - $k = k + 1$
- while  $|f(x^{k+1}) - f(x^k)| > \varepsilon$

# Convergence of Gradient Descent



- For appropriate step size, for convex (bowl-shaped) functions gradient descent will always find the minimum.



- For non-convex functions it will find a local minimum or an inflection point

- Returning to our problem..

# Problem Statement

- Given a training set of input-output pairs  $(X_1, \mathbf{d}_1), (X_2, \mathbf{d}_2), \dots, (X_T, \mathbf{d}_T)$
- Minimize the following function

$$Loss(W) = \frac{1}{T} \sum_i div(f(X_i; W), d_i)$$

w.r.t  $W$

- This is problem of function minimization
  - An instance of optimization

# Preliminaries

- Before we proceed: the problem setup

# Problem Setup: Things to define

- Given a training set of input-output pairs  
 $(X_1, d_1), (X_2, d_2), \dots, (X_T, d_T)$
- What are these input-output pairs?

$$Loss(W) = \frac{1}{T} \sum_i div(f(X_i; W), d_i)$$

# Problem Setup: Things to define

- Given a training set of input-output pairs  $(X_1, d_1), (X_2, d_2), \dots, (X_T, d_T)$
- What are these input-output pairs?

$$Loss(W) = \frac{1}{T} \sum_i div(f(X_i; W), d_i)$$

What is  $f()$  and  
what are its  
parameters  $W$ ?

# Problem Setup: Things to define

- Given a training set of input-output pairs  $(X_1, d_1), (X_2, d_2), \dots, (X_T, d_T)$
- What are these input-output pairs?

$$Loss(W) = \frac{1}{T} \sum_i div(f(X_i; W), d_i)$$

What is the divergence  $div()$ ?

What is  $f()$  and what are its parameters  $W$ ?

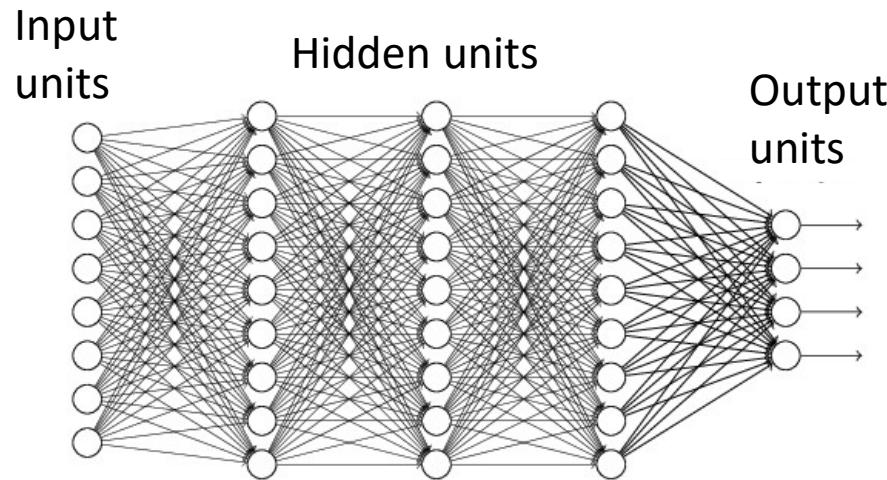
# Problem Setup: Things to define

- Given a training set of input-output pairs  $(X_1, d_1), (X_2, d_2), \dots, (X_T, d_T)$
- Minimize the following function

$$Loss(W) = \frac{1}{T} \sum_i div(f(X_i; W), d_i)$$

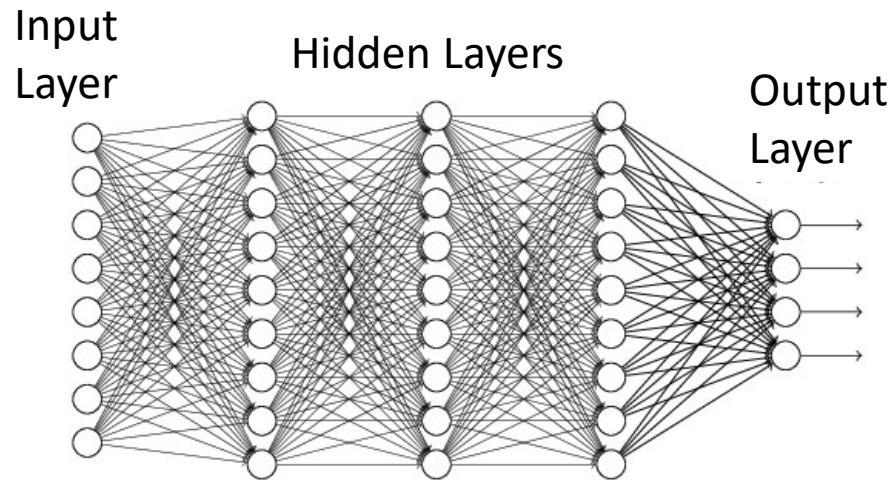
What is  $f()$  and  
what are its  
parameters  $W$ ?

# What is $f()$ ? Typical network



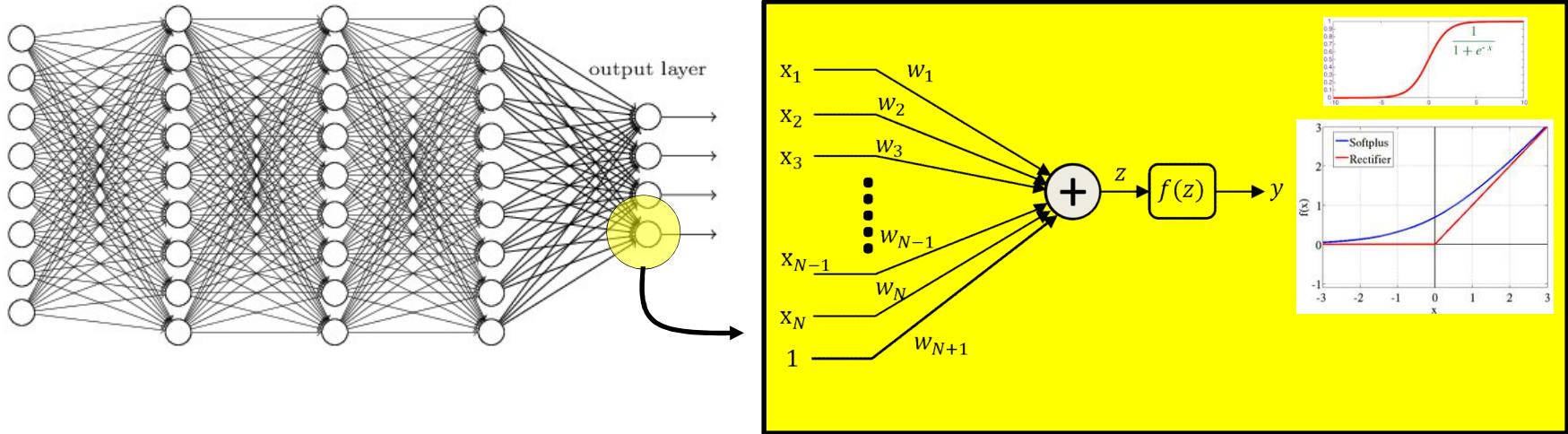
- Multi-layer perceptron
- A *directed* network with a set of inputs and outputs
  - No loops

# Typical network



- We assume a “layered” network for simplicity
  - Each “layer” of neurons only gets inputs from the earlier layer(s) and outputs signals only to later layer(s)
  - We will refer to the inputs as the ***input layer***
    - No neurons here – the “layer” simply refers to inputs
  - We refer to the outputs as the ***output layer***
  - Intermediate layers are ***“hidden” layers***

# The individual neurons



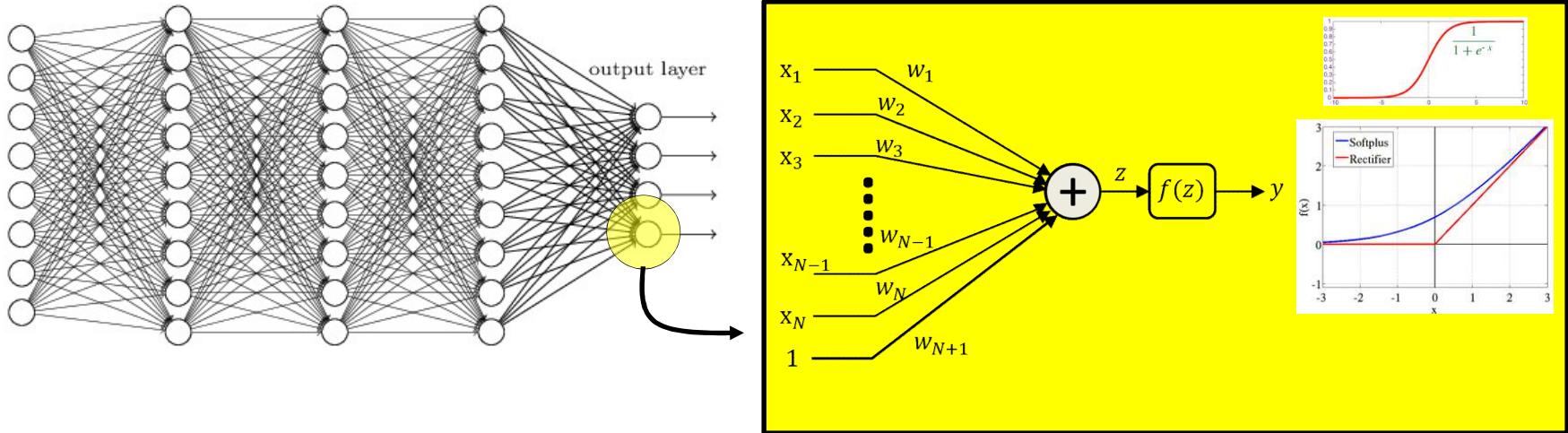
- Individual neurons operate on a set of inputs and produce a single output
  - **Standard setup:** A differentiable activation function applied to an affine combination of the inputs

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

- More generally: *any* differentiable function

$$y = f(x_1, x_2, \dots, x_N; W)$$

# The individual neurons



- Individual neurons operate on a set of inputs and produce a single output

- **Standard setup:** A differentiable activation function applied to an affine combination of the input

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

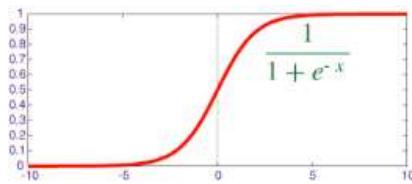
- More generally: *any* differentiable function

$$y = f(x_1, x_2, \dots, x_N; W)$$

We will assume this unless otherwise specified

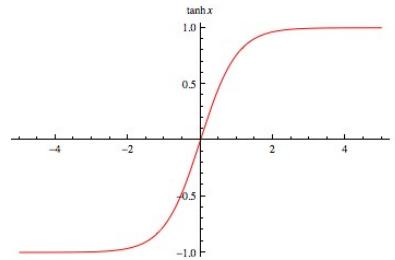
Parameters are weights  $w_i$  and bias  $b$

# Activations and their derivatives



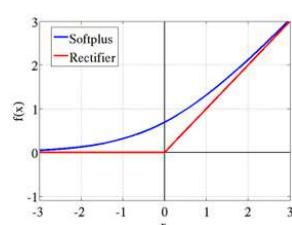
$$f(z) = \frac{1}{1 + \exp(-z)}$$

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



$$f(z) = \tanh(z)$$

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



$$f(z) = \begin{cases} z, & z \geq 0 \\ 0, & z < 0 \end{cases}$$

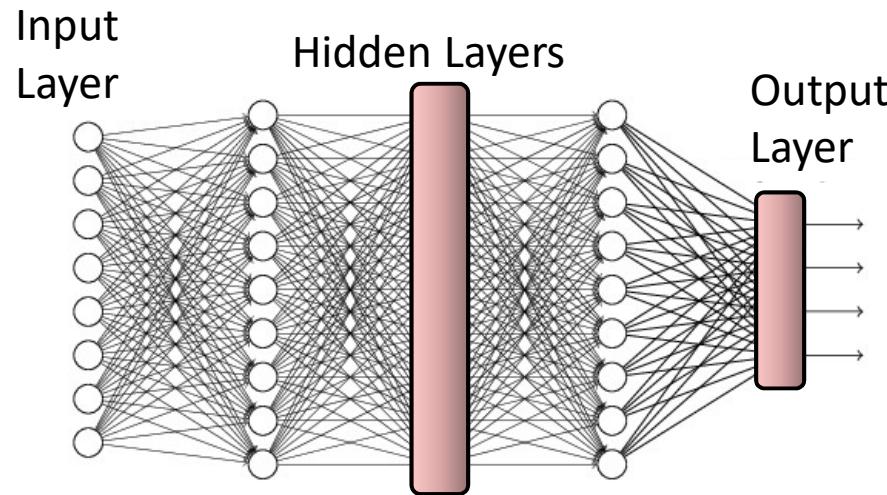
[\*]  $f'(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases}$

$$f(z) = \log(1 + \exp(z))$$

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

- Some popular activation functions and their derivatives

# Vector Activations

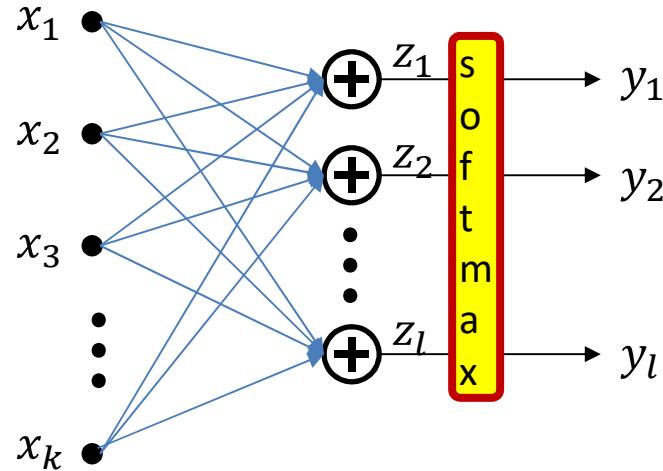


- We can also have neurons that have *multiple coupled* outputs

$$[y_1, y_2, \dots, y_l] = f(x_1, x_2, \dots, x_k; W)$$

- Function  $f()$  operates on set of inputs to produce set of outputs
- Modifying a single parameter in  $W$  will affect *all* outputs

# Vector activation example: Softmax



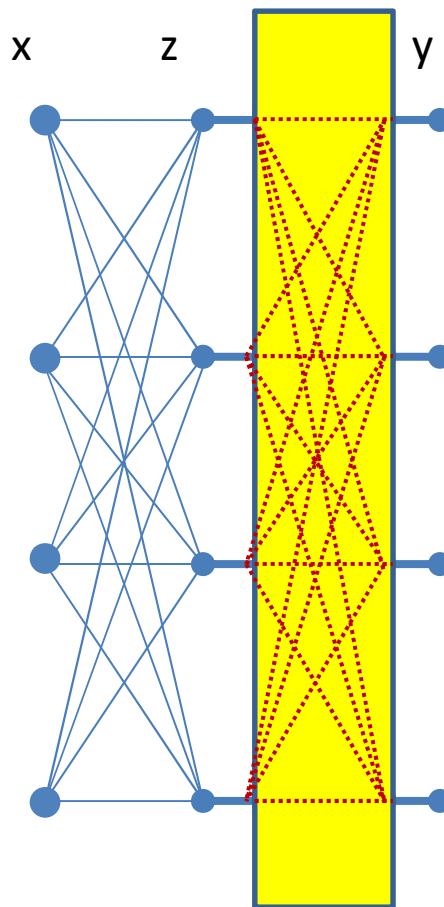
- Example: Softmax *vector* activation

$$z_i = \sum_j w_{ji} x_j + b_i$$

$$y = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

Parameters are  
weights  $w_{ji}$   
and bias  $b_i$

# Multiplicative combination: Can be viewed as a case of vector activations



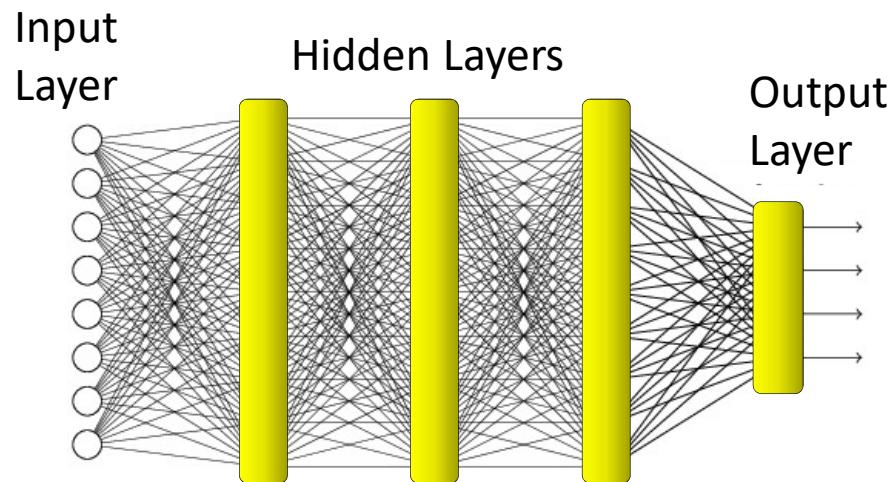
$$z_i = \sum_j w_{ji} x_j + b_i$$

$$y_i = \prod_l (z_l)^{\alpha_{li}}$$

Parameters are  
weights  $w_{ji}$   
and bias  $b_i$

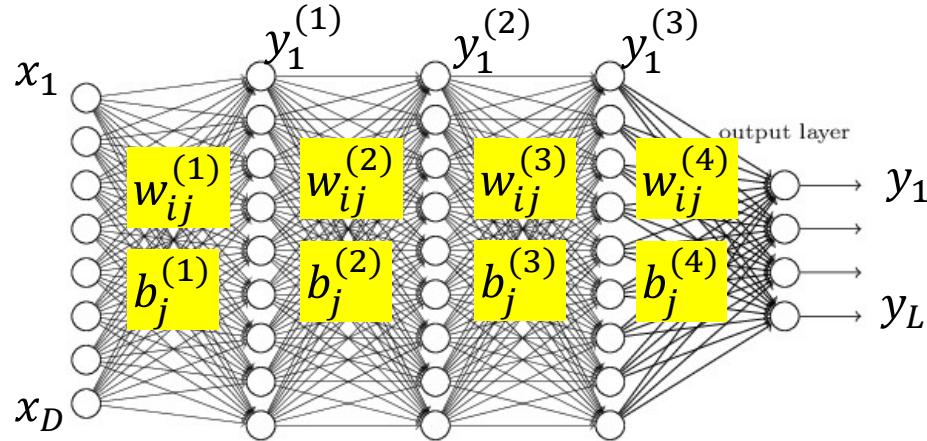
- A layer of multiplicative combination is a special case of vector activation

# Typical network



- In a layered network, each layer of perceptrons can be viewed as a single vector activation

# Notation



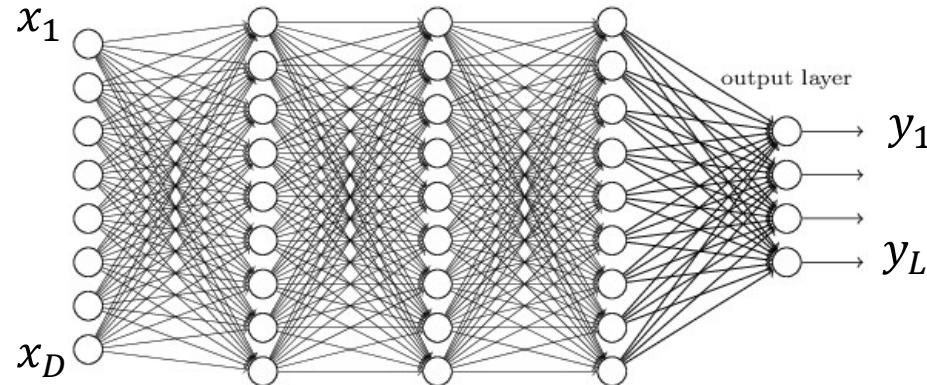
- The input layer is the  $0^{\text{th}}$  layer
- We will represent the output of the  $i$ -th perceptron of the  $k^{\text{th}}$  layer as  $y_i^{(k)}$ 
  - **Input to network:**  $y_i^{(0)} = x_i$
  - **Output of network:**  $y_i = y_i^{(N)}$
- We will represent the weight of the connection between the  $i$ -th unit of the  $k-1$ th layer and the  $j$ th unit of the  $k$ -th layer as  $w_{ij}^{(k)}$ 
  - The bias to the  $j$ th unit of the  $k$ -th layer is  $b_j^{(k)}$

# Problem Setup: Things to define

- Given a training set of input-output pairs  
 $(X_1, d_1), (X_2, d_2), \dots, (X_T, d_T)$
- What are these input-output pairs?

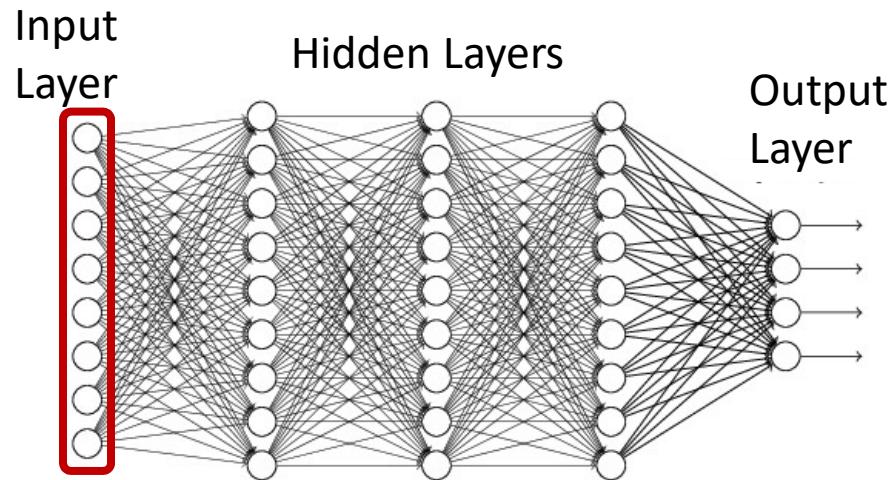
$$Loss(W) = \frac{1}{T} \sum_i div(f(X_i; W), d_i)$$

# Vector notation



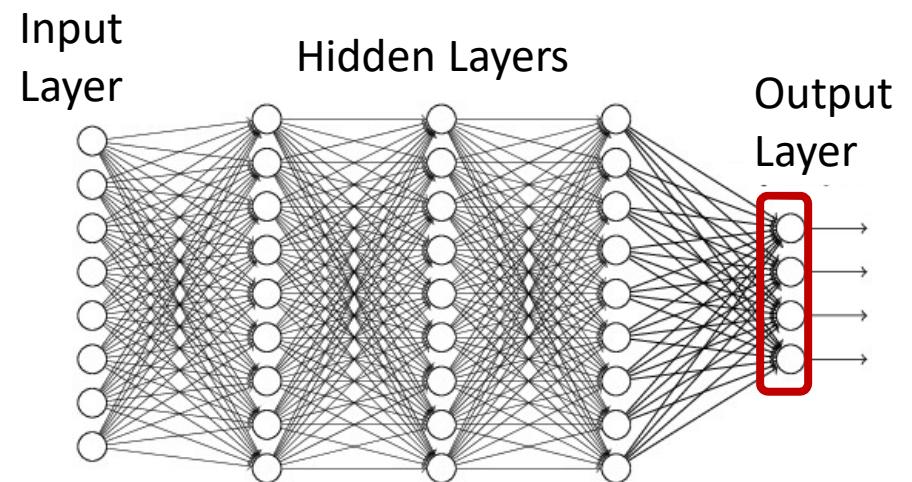
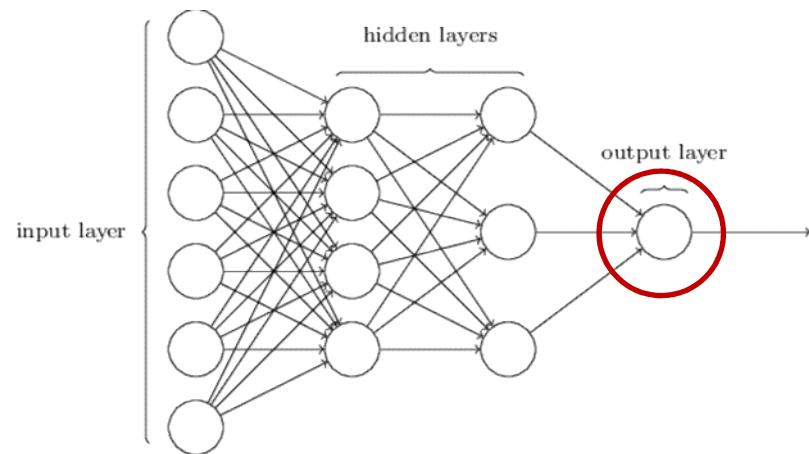
- Given a training set of input-output pairs  $(X_1, d_1), (X_2, d_2), \dots, (X_T, d_T)$
- $X_n = [x_{n1}, x_{n2}, \dots, x_{nD}]$  is the nth input vector
- $d_n = [d_{n1}, d_{n2}, \dots, d_{nL}]$  is the nth desired output
- $Y_n = [y_{n1}, y_{n2}, \dots, y_{nL}]$  is the nth vector of *actual* outputs of the network
- We will sometimes drop the first subscript when referring to a *specific* instance

# Representing the input



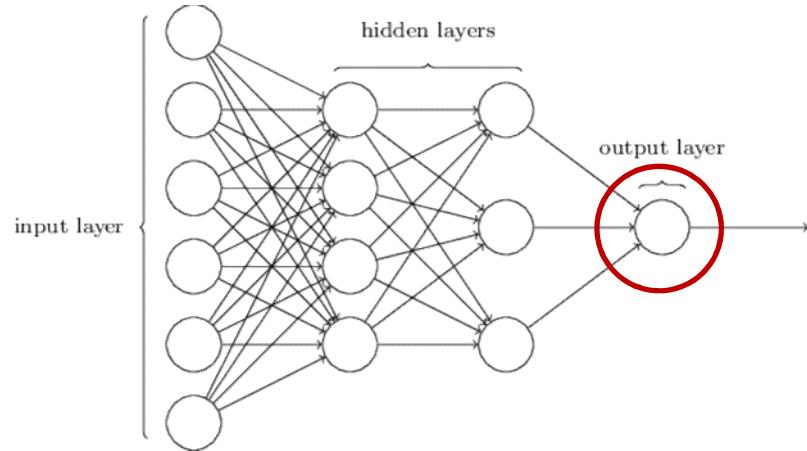
- Vectors of numbers
  - (or may even be just a scalar, if input layer is of size 1)
  - E.g. vector of pixel values
  - E.g. vector of speech features
  - E.g. real-valued vector representing text
    - We will see how this happens later in the course
  - Other real valued vectors

# Representing the output



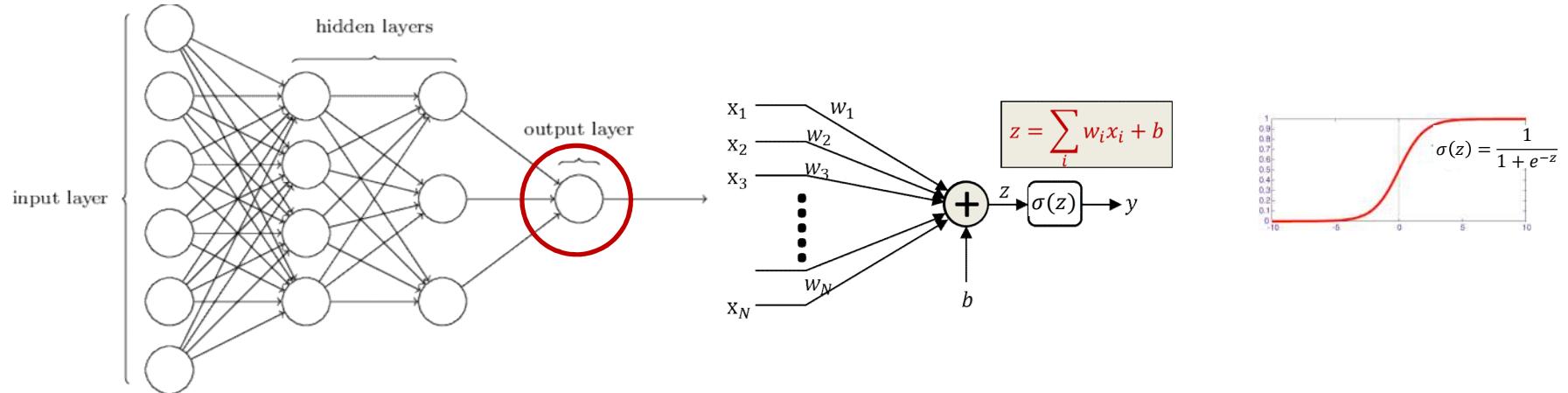
- If the desired *output* is real-valued, no special tricks are necessary
  - Scalar Output : single output neuron
    - $d = \text{scalar} (\text{real value})$
  - Vector Output : as many output neurons as the dimension of the desired output
    - $d = [d_1 \ d_2 \dots \ d_L]$  (vector of real values)

# Representing the output



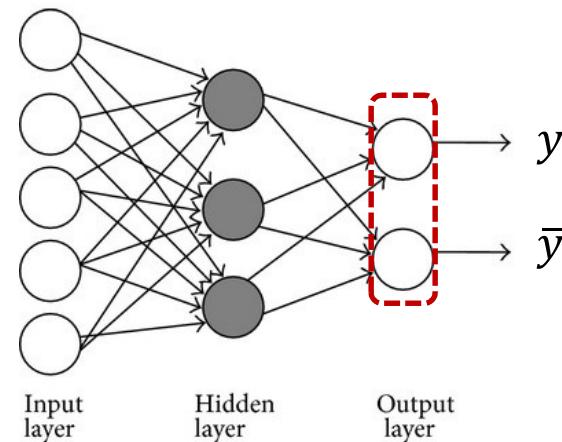
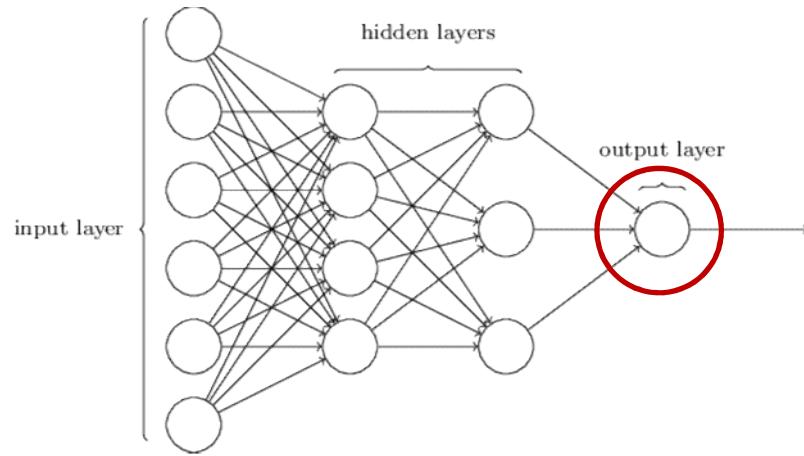
- If the desired output is binary (is this a cat or not), use a simple 1/0 representation of the desired output
  - 1 = Yes it's a cat
  - 0 = No it's not a cat.

# Representing the output



- If the desired output is binary (is this a cat or not), use a simple 1/0 representation of the desired output
- Output activation: Typically a sigmoid
  - Viewed as the *probability*  $P(Y = 1|X)$  of class value 1
    - Indicating the fact that for actual data, in general a feature value X may occur for both classes, but with different probabilities
    - Is differentiable

# Representing the output

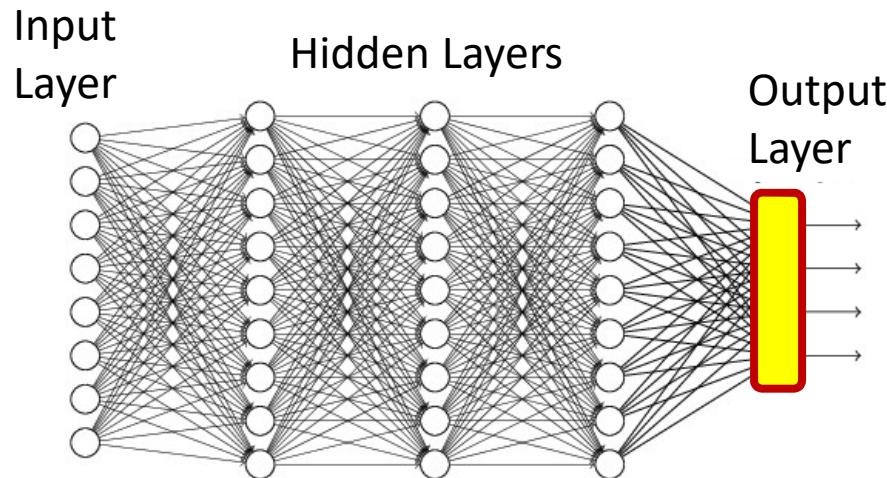


- If the desired output is binary (is this a cat or not), use a simple 1/0 representation of the desired output
  - 1 = Yes it's a cat
  - 0 = No it's not a cat.
- Sometimes represented by *two* outputs, one representing the desired output, the other representing the *negation* of the desired output
  - Yes:  $\rightarrow [1 \ 0]$
  - No:  $\rightarrow [0 \ 1]$
- The output explicitly becomes a 2-output softmax

# Multi-class output: One-hot representations

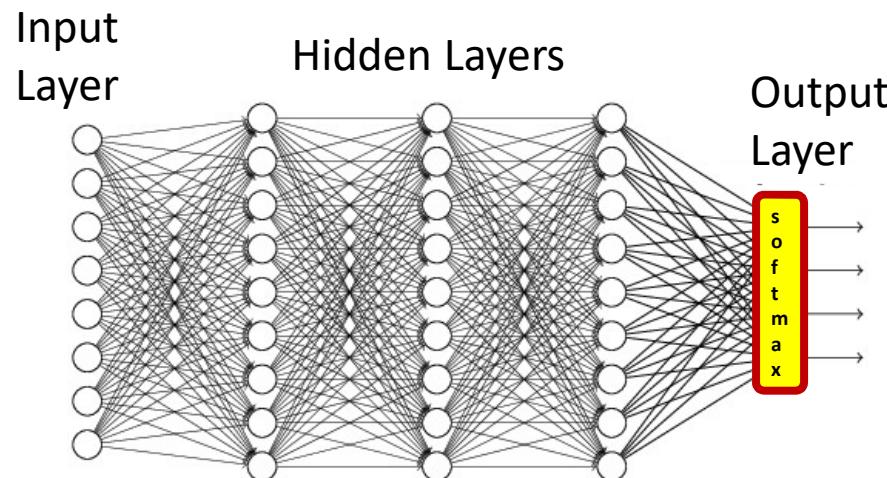
- Consider a network that must distinguish if an input is a cat, a dog, a camel, a hat, or a flower
- We can represent this set as the following vector:
$$[\text{cat } \text{dog } \text{camel } \text{hat } \text{flower}]^T$$
- For inputs of each of the five classes the desired output is:
  - cat:  $[1 \ 0 \ 0 \ 0 \ 0]^T$
  - dog:  $[0 \ 1 \ 0 \ 0 \ 0]^T$
  - camel:  $[0 \ 0 \ 1 \ 0 \ 0]^T$
  - hat:  $[0 \ 0 \ 0 \ 1 \ 0]^T$
  - flower:  $[0 \ 0 \ 0 \ 0 \ 1]^T$
- For an input of any class, we will have a five-dimensional vector output with four zeros and a single 1 at the position of that class
- This is a *one hot vector*

# Multi-class networks



- For a multi-class classifier with  $N$  classes, the one-hot representation will have  $N$  binary target outputs ( $d$ )
  - An  $N$ -dimensional binary vector
- The neural network's output too must ideally be binary ( $N-1$  zeros and a single 1 in the right place)
- More realistically, it will be a probability vector
  - $N$  probability values that sum to 1.

# Multi-class classification: Output



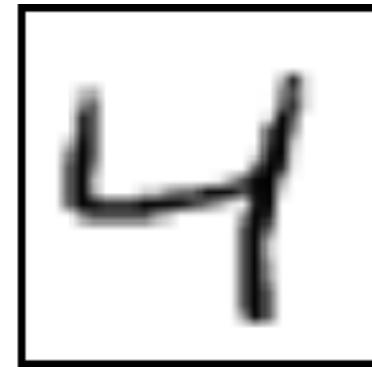
- Softmax *vector* activation is often used at the output of multi-class classifier nets

$$z_i = \sum_j w_{ji}^{(n)} y_j^{(n-1)}$$

$$y_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

- This can be viewed as the probability  $y_i = P(\text{class} = i | X)$

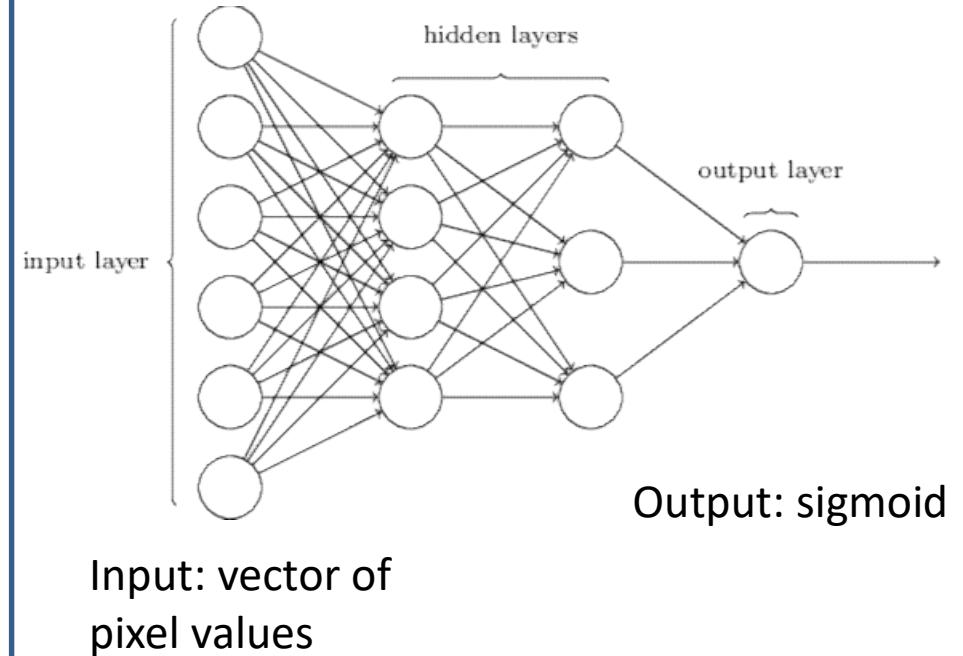
# Typical Problem Statement



- We are given a number of “training” data instances
- E.g. images of digits, along with information about which digit the image represents
- Tasks:
  - Binary recognition: Is this a “2” or not
  - Multi-class recognition: Which digit is this? Is this a digit in the first place?

# Typical Problem statement: binary classification

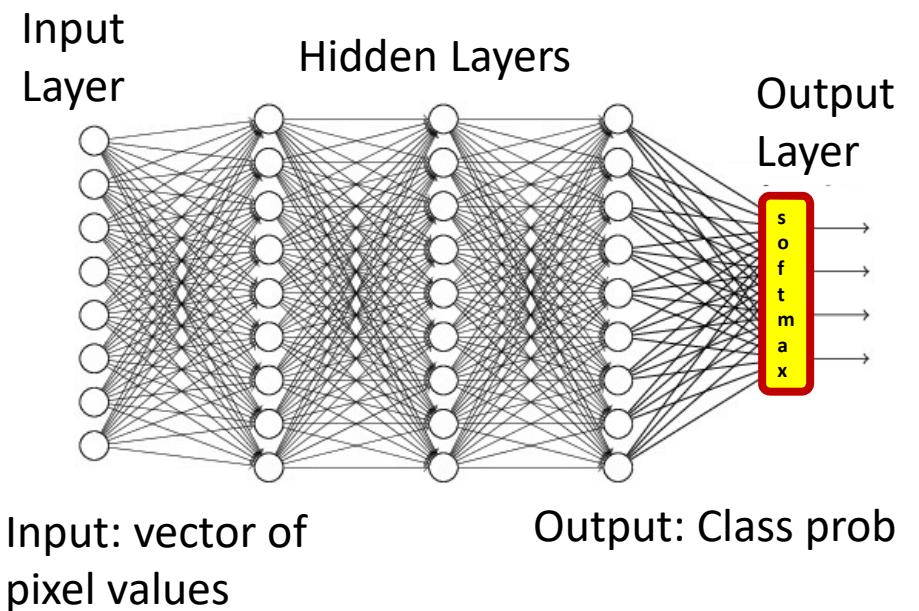
| Training data |        |
|---------------|--------|
| (Σ, 0)        | (2, 1) |
| (2, 1)        | (4, 0) |
| (0, 0)        | (2, 1) |



- Given, many positive and negative examples (training data),
  - learn all weights such that the network does the desired job

# Typical Problem statement: multiclass classification

| Training data |        |
|---------------|--------|
| (Σ, 5)        | (2, 2) |
| (2, 2)        | (4, 4) |
| (0, 0)        | (2, 2) |



- Given, many positive and negative examples (training data),
  - learn all weights such that the network does the desired job

# Problem Setup: Things to define

- Given a training set of input-output pairs  $(X_1, d_1), (X_2, d_2), \dots, (X_T, d_T)$
- Minimize the following function

$$Loss(W) = \frac{1}{T} \sum_i div(f(X_i; W), d_i)$$

What is the  
divergence  $div()$ ?

# Problem Setup: Things to define

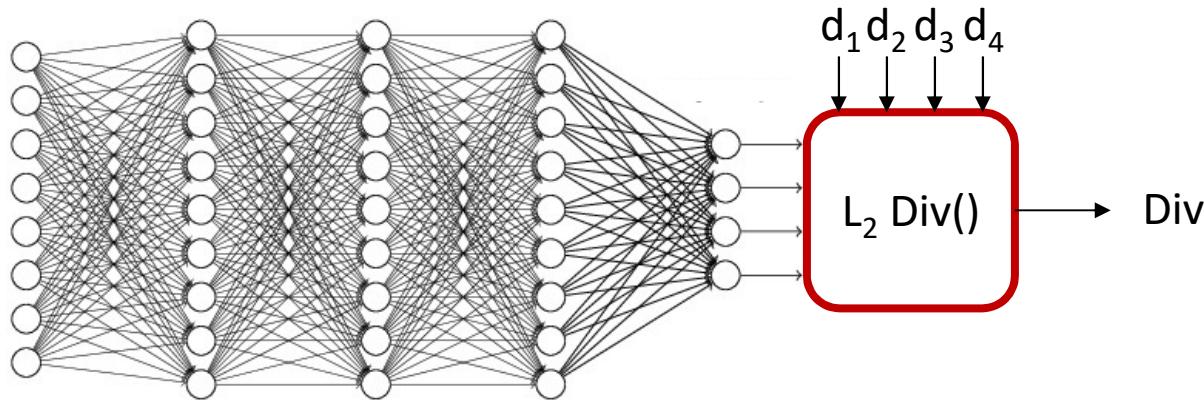
- Given a training set of input-output pairs  $(X_1, d_1), (X_2, d_2), \dots, (X_T, d_T)$
- Minimize the following function

$$Loss(W) = \frac{1}{T} \sum_i div(f(X_i; W), d_i)$$

What is the divergence  $div()$ ?

Note: For  $Loss(W)$  to be differentiable w.r.t  $W$ ,  $div()$  must be differentiable

# Examples of divergence functions



- For real-valued output vectors, the (scaled) L<sub>2</sub> divergence is popular

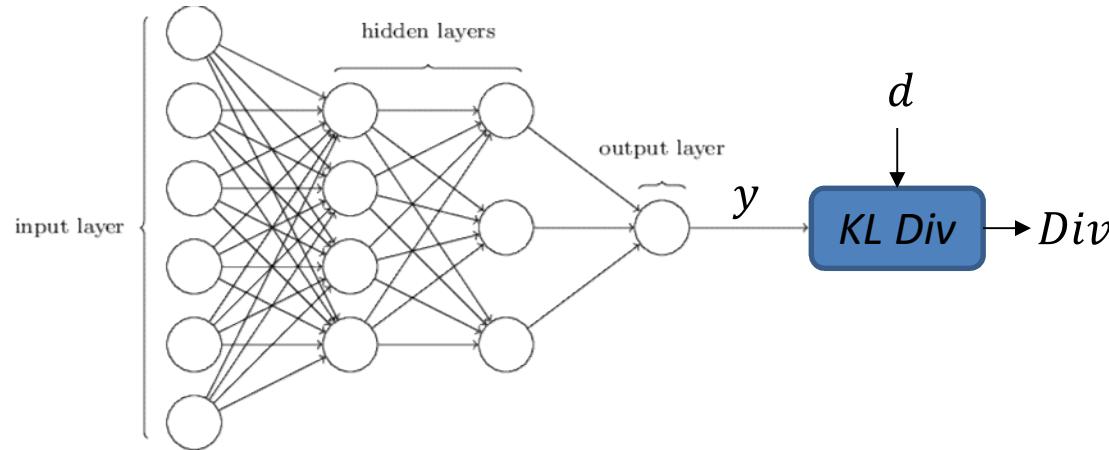
$$Div(Y, d) = \frac{1}{2} \|Y - d\|^2 = \frac{1}{2} \sum_i (y_i - d_i)^2$$

- Squared Euclidean distance between true and desired output
- Note: this is differentiable

$$\frac{dDiv(Y, d)}{dy_i} = (y_i - d_i)$$

$$\nabla_Y Div(Y, d) = [y_1 - d_1, y_2 - d_2, \dots]$$

# For binary classifier



- For binary classifier with scalar output,  $Y \in (0,1)$ ,  $d$  is 0/1, the cross entropy between the probability distribution  $[Y, 1 - Y]$  and the ideal output probability  $[d, 1 - d]$  is popular

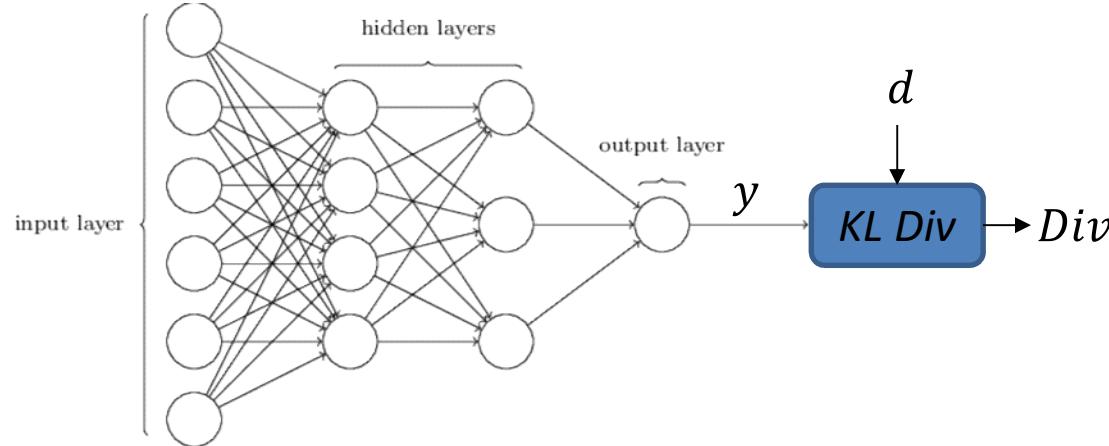
$$Div(Y, d) = -d\log Y - (1 - d)\log(1 - Y)$$

- Minimum when  $d = Y$

- Derivative

$$\frac{dDiv(Y, d)}{dY} = \begin{cases} -\frac{1}{Y} & \text{if } d = 1 \\ \frac{1}{1 - Y} & \text{if } d = 0 \end{cases}$$

# For binary classifier



- For binary classifier with scalar output,  $Y \in (0,1)$ ,  $d$  is 0/1, the cross entropy between the probability distribution  $[Y, 1 - Y]$  and the ideal output probability  $[d, 1 - d]$  is popular

$$Div(Y, d) = -d\log Y - (1 - d)\log(1 - Y)$$

- Minimum when  $d = Y$

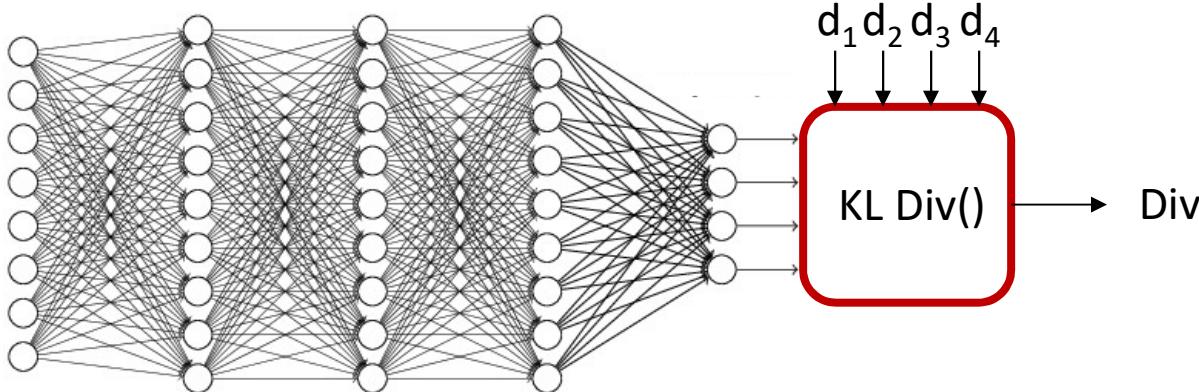
- Derivative

$$\frac{dDiv(Y, d)}{dY} = \begin{cases} -\frac{1}{Y} & \text{if } d = 1 \\ \frac{1}{1 - Y} & \text{if } d = 0 \end{cases}$$

Note: when  $y = d$  the derivative is *not* 0

*Even though div() = 0  
(minimum) when  $y = d$*

# For multi-class classification



- Desired output  $d$  is a one hot vector  $[0 \ 0 \dots 1 \ \dots 0 \ 0 \ 0]$  with the 1 in the  $c$ -th position (for class  $c$ )
- Actual output will be probability distribution  $[y_1, y_2, \dots]$
- The cross-entropy between the desired one-hot output and actual output:

$$Div(Y, d) = - \sum_i d_i \log y_i = - \log y_c$$

- Derivative

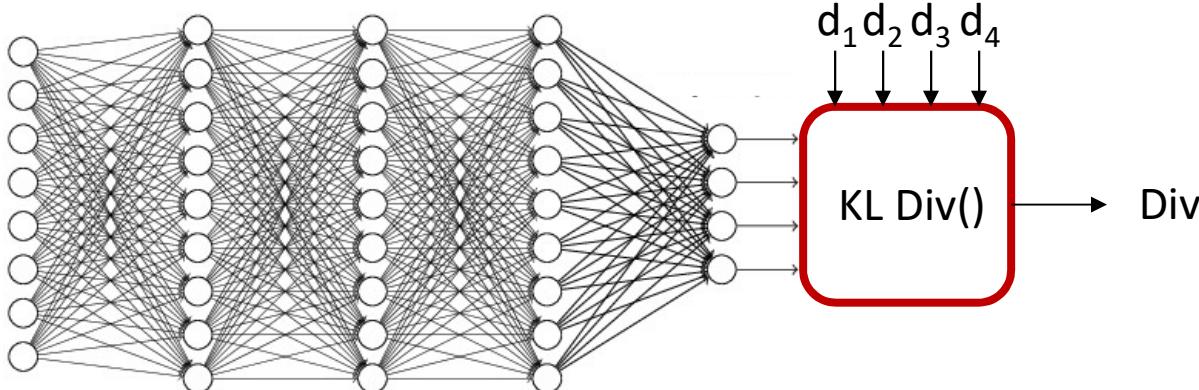
$$\frac{dDiv(Y, d)}{dY_i} = \begin{cases} -\frac{1}{y_c} & \text{for the } c-\text{th component} \\ 0 & \text{for remaining component} \end{cases}$$

$$\nabla_Y Div(Y, d) = \left[ 0 \ 0 \ \dots \frac{-1}{y_c} \dots 0 \ 0 \right]$$

If  $y_c < 1$ , the slope is negative w.r.t.  $y_c$

Indicates *increasing*  $y_c$  will *reduce* divergence

# For multi-class classification



- Desired output  $d$  is a one hot vector  $[0 \ 0 \dots 1 \ \dots 0 \ 0 \ 0]$  with the 1 in the  $c$ -th position (for class  $c$ )
- Actual output will be probability distribution  $[y_1, y_2, \dots]$
- The cross-entropy between the desired one-hot output and actual output:

$$Div(Y, d) = - \sum_i d_i \log y_i = - \log y_c$$

- Derivative

$$\frac{dDiv(Y, d)}{dY_i} = \begin{cases} -\frac{1}{y_c} & \text{for the } c-\text{th component} \\ 0 & \text{for remaining component} \end{cases}$$

$$\nabla_Y Div(Y, d) = \left[ 0 \ 0 \ \dots \frac{-1}{y_c} \dots \ 0 \ 0 \right]$$

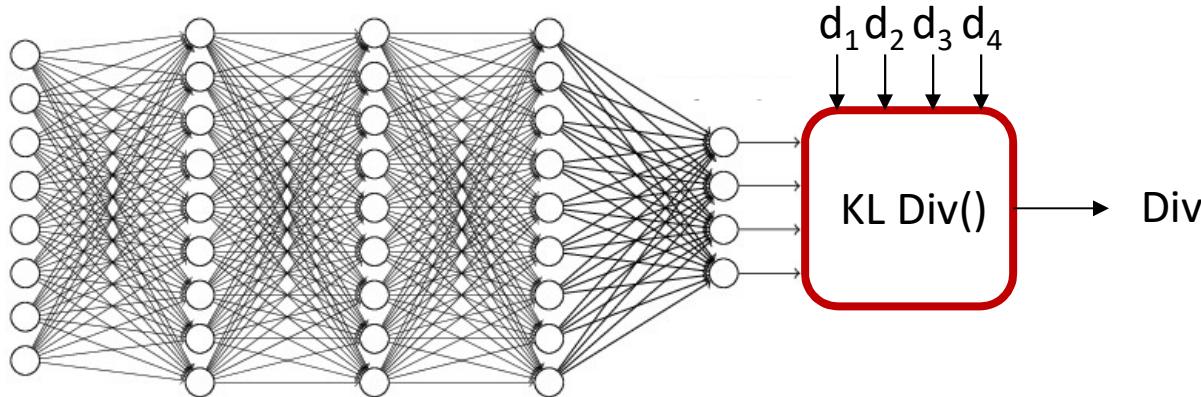
If  $y_c < 1$ , the slope is negative w.r.t.  $y_c$

Indicates *increasing*  $y_c$  will *reduce* divergence

Note: when  $y = d$  the derivative is *not* 0

*Even though div() = 0 (minimum) when  $y = d$*

# For multi-class classification



- It is sometimes useful to set the target output to  $[\epsilon \ \epsilon \dots (1 - (K - 1)\epsilon) \dots \epsilon \ \epsilon \ \epsilon]$  with the value  $1 - (K - 1)\epsilon$  in the  $c$ -th position (for class  $c$ ) and  $\epsilon$  elsewhere for some small  $\epsilon$ 
  - “Label smoothing” -- aids gradient descent
- The cross-entropy remains:

$$Div(Y, d) = - \sum_i d_i \log y_i$$

- Derivative

$$\frac{dDiv(Y, d)}{dY_i} = \begin{cases} -\frac{1 - (K - 1)\epsilon}{y_c} & \text{for the } c\text{-th component} \\ -\frac{\epsilon}{y_i} & \text{for remaining components} \end{cases}$$

# Problem Setup: Things to define

- Given a training set of input-output pairs  $(X_1, d_1), (X_2, d_2), \dots, (X_T, d_T)$
- Minimize the following function

$$Loss(W) = \frac{1}{T} \sum_i div(f(X_i; W), d_i)$$

ALL TERMS HAVE BEEN DEFINED

# Problem Setup

- Given a training set of input-output pairs  $(X_1, d_1), (X_2, d_2), \dots, (X_T, d_T)$
- The error on the  $i^{\text{th}}$  instance is  $\text{div}(Y_i, d_i)$ 
  - $Y_i = f(X_i; W)$
- The loss

$$\textit{Loss} = \frac{1}{T} \sum_i \text{div}(Y_i, d_i)$$

- Minimize  $\textit{Loss}$  w.r.t  $\{w_{ij}^{(k)}, b_j^{(k)}\}$

# Recap: Gradient Descent Algorithm

- Initialize:

- $x^0$

- $k = 0$

To minimize any function  $f(x)$  w.r.t  $x$

- do

- $x^{k+1} = x^k - \eta^k \nabla f(x^k)^T$

- $k = k + 1$

- while  $|f(x^k) - f(x^{k-1})| > \varepsilon$

# Recap: Gradient Descent Algorithm

- In order to minimize any function  $f(x)$  w.r.t.  $x$
- Initialize:
  - $x^0$
  - $k = 0$
- do
  - For every component  $i$ 
    - $x_i^{k+1} = x_i^k - \eta^k \frac{\partial f}{\partial x_i}$  Explicitly stating it by component
  - $k = k + 1$
- while  $|f(x^k) - f(x^{k-1})| > \varepsilon$

# Training Neural Nets through Gradient Descent

Total training Loss:

$$Loss = \frac{1}{T} \sum_t Div(\mathbf{Y}_t, \mathbf{d}_t)$$

- Gradient descent algorithm:
- Initialize all weights and biases  $\{w_{ij}^{(k)}\}$ 
  - Using the extended notation: the bias is also a weight
- Do:
  - For every layer  $k$  for all  $i, j$ , update:
    - $w_{i,j}^{(k)} = w_{i,j}^{(k)} - \eta \frac{dLoss}{dw_{i,j}^{(k)}}$
- Until  $Loss$  has converged

Assuming the bias is also represented as a weight

# Training Neural Nets through Gradient Descent

Total training Loss:

$$Loss = \frac{1}{T} \sum_t Div(\mathbf{Y}_t, \mathbf{d}_t)$$

- Gradient descent algorithm:
- Initialize all weights  $\{w_{ij}^{(k)}\}$
- Do:
  - For every layer  $k$  for all  $i, j$ , update:
    - $w_{i,j}^{(k)} = w_{i,j}^{(k)} - \eta \frac{dLoss}{dw_{i,j}^{(k)}}$
- Until  $Err$  has converged

# The derivative

Total training Loss:

$$Loss = \frac{1}{T} \sum_t Div(\mathbf{Y}_t, \mathbf{d}_t)$$

- Computing the derivative

Total derivative:

$$\frac{dLoss}{dw_{i,j}^{(k)}} = \frac{1}{T} \sum_t \frac{dDiv(\mathbf{Y}_t, \mathbf{d}_t)}{dw_{i,j}^{(k)}}$$

# Training by gradient descent

- Initialize all weights  $\{w_{ij}^{(k)}\}$
- Do:
  - For all  $i, j, k$ , initialize  $\frac{dLoss}{dw_{i,j}^{(k)}} = 0$
  - For all  $t = 1:T$ 
    - For every layer  $k$  for all  $i, j$ :
      - Compute  $\frac{dDiv(\mathbf{Y}_t, \mathbf{d}_t)}{dw_{i,j}^{(k)}}$
      - $\frac{dLoss}{dw_{i,j}^{(k)}} += \frac{dDiv(\mathbf{Y}_t, \mathbf{d}_t)}{dw_{i,j}^{(k)}}$
    - For every layer  $k$  for all  $i, j$ :
$$w_{i,j}^{(k)} = w_{i,j}^{(k)} - \frac{\eta}{T} \frac{dLoss}{dw_{i,j}^{(k)}}$$
  - Until  $Err$  has converged

# The derivative

Total training Loss:

$$Loss = \frac{1}{T} \sum_t Div(\mathbf{Y}_t, \mathbf{d}_t)$$

Total derivative:

$$\frac{dLoss}{dw_{i,j}^{(k)}} = \frac{1}{T} \sum_t \frac{dDiv(\mathbf{Y}_t, \mathbf{d}_t)}{dw_{i,j}^{(k)}}$$

- So we must first figure out how to compute the derivative of divergences of individual training inputs

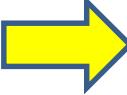
# Calculus Refresher: Basic rules of calculus

For any differentiable function

$$y = f(x)$$

with derivative

$$\frac{dy}{dx}$$

the following must hold for sufficiently small  $\Delta x$    $\Delta y \approx \frac{dy}{dx} \Delta x$

For any differentiable function

$$y = f(x_1, x_2, \dots, x_M)$$

with partial derivatives

$$\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \dots, \frac{\partial y}{\partial x_M}$$

the following must hold for sufficiently small  $\Delta x_1, \Delta x_2, \dots, \Delta x_M$

Both by the definition  
 $\Delta y = \nabla f \Delta x$

$$\Delta y \approx \frac{\partial y}{\partial x_1} \Delta x_1 + \frac{\partial y}{\partial x_2} \Delta x_2 + \dots + \frac{\partial y}{\partial x_M} \Delta x_M$$

# Calculus Refresher: Chain rule

For any nested function  $y = f(g(x))$

$$\frac{dy}{dx} = \frac{\partial f}{\partial g(x)} \frac{dg(x)}{dx}$$

Check - we can confirm that :  $\Delta y = \frac{dy}{dx} \Delta x$

$$z = g(x) \rightarrow \Delta z = \frac{dg(x)}{dx} \Delta x$$

$$y = f(z) \rightarrow \Delta y = \frac{df}{dz} \Delta z = \frac{df}{dz} \frac{dg(x)}{dx} \Delta x$$



# Calculus Refresher: Distributed Chain rule

$$y = f(g_1(x), g_2(x), \dots, g_M(x))$$

$$\frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$$

Check:  $\Delta y = \frac{dy}{dx} \Delta x$       Let  $z_i = g_i(x)$

$$\Delta y = \frac{\partial f}{\partial z_1} \Delta z_1 + \frac{\partial f}{\partial z_2} \Delta z_2 + \dots + \frac{\partial f}{\partial z_M} \Delta z_M$$

$$\Delta y = \frac{\partial f}{\partial z_1} \frac{dz_1}{dx} \Delta x + \frac{\partial f}{\partial z_2} \frac{dz_2}{dx} \Delta x + \dots + \frac{\partial f}{\partial z_M} \frac{dz_M}{dx} \Delta x$$

$$\Delta y = \left( \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx} \right) \Delta x$$



# Calculus Refresher: Distributed Chain rule

$$y = f(g_1(x), g_2(x), \dots, g_M(x))$$

$$\frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$$

Check:  $\Delta y = \frac{dy}{dx} \Delta x$

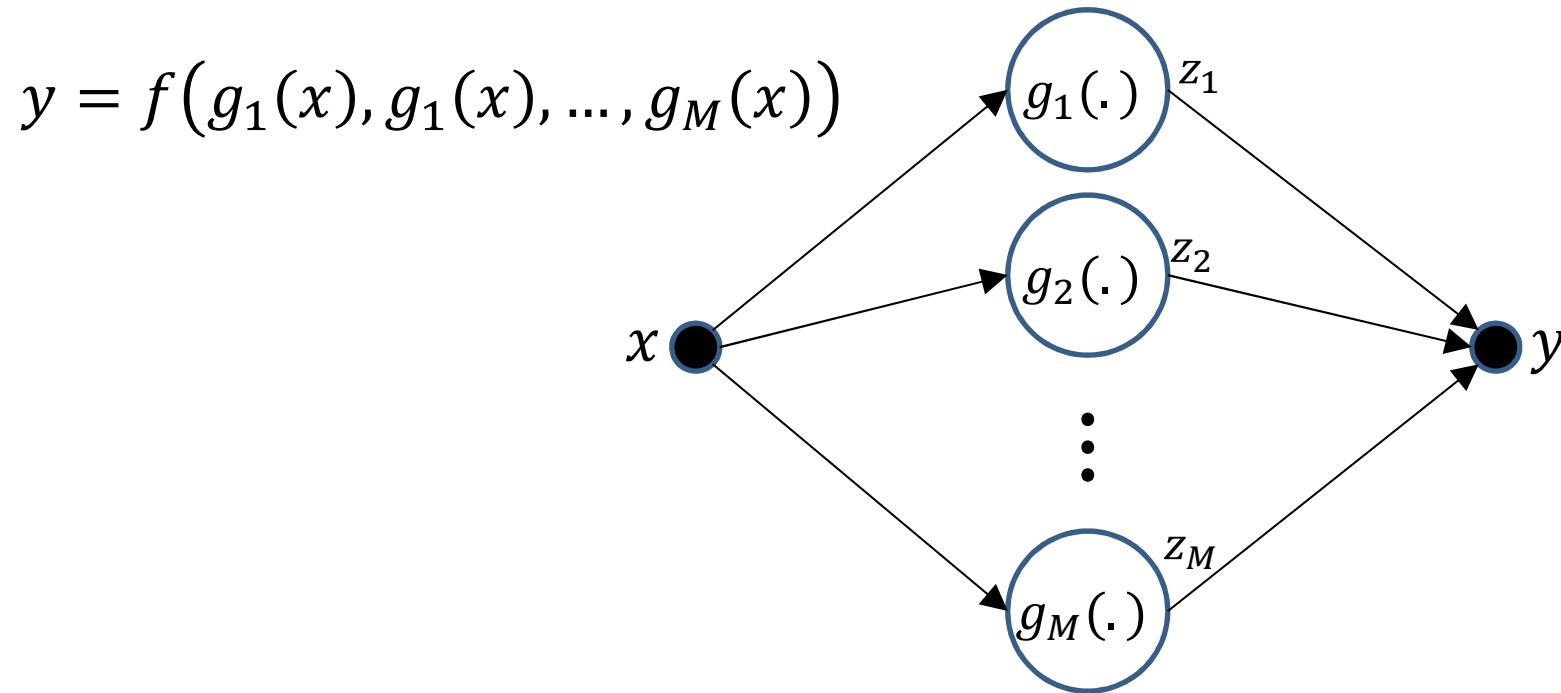
$$\Delta y = \frac{\partial f}{\partial g_1(x)} \Delta g_1(x) + \frac{\partial f}{\partial g_2(x)} \Delta g_2(x) + \dots + \frac{\partial f}{\partial g_M(x)} \Delta g_M(x)$$

$$\Delta y = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} \Delta x + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} \Delta x + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx} \Delta x$$

$$\Delta y = \left( \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \dots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx} \right) \Delta x$$

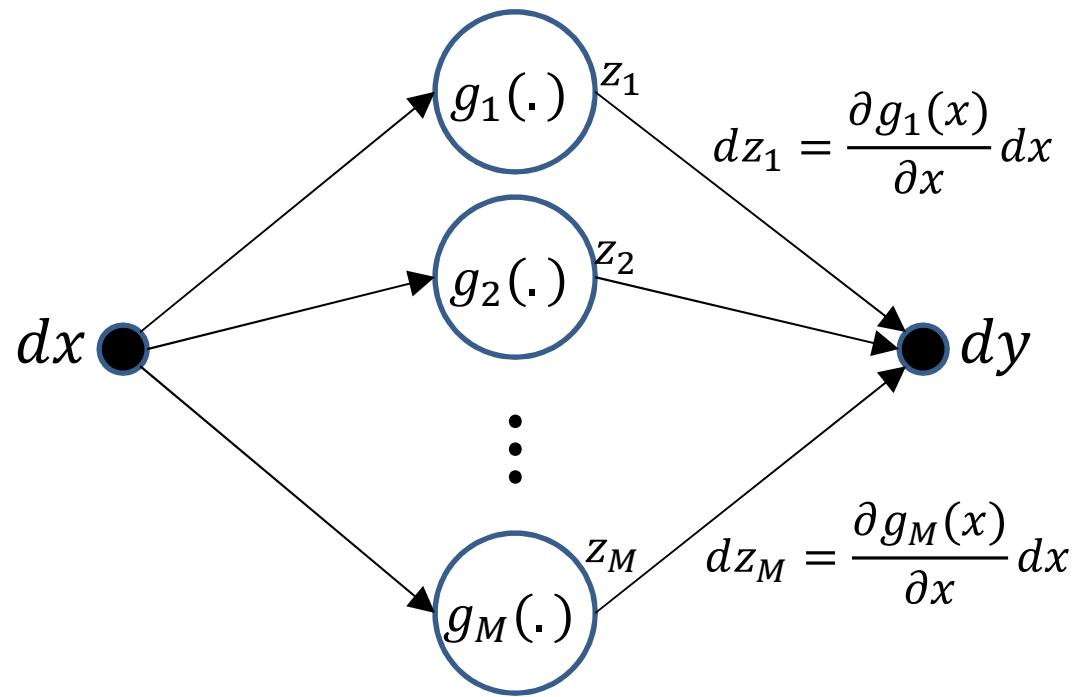


# Distributed Chain Rule: Influence Diagram



- $x$  affects  $y$  through each of  $g_1 \dots g_M$

# Distributed Chain Rule: Influence Diagram

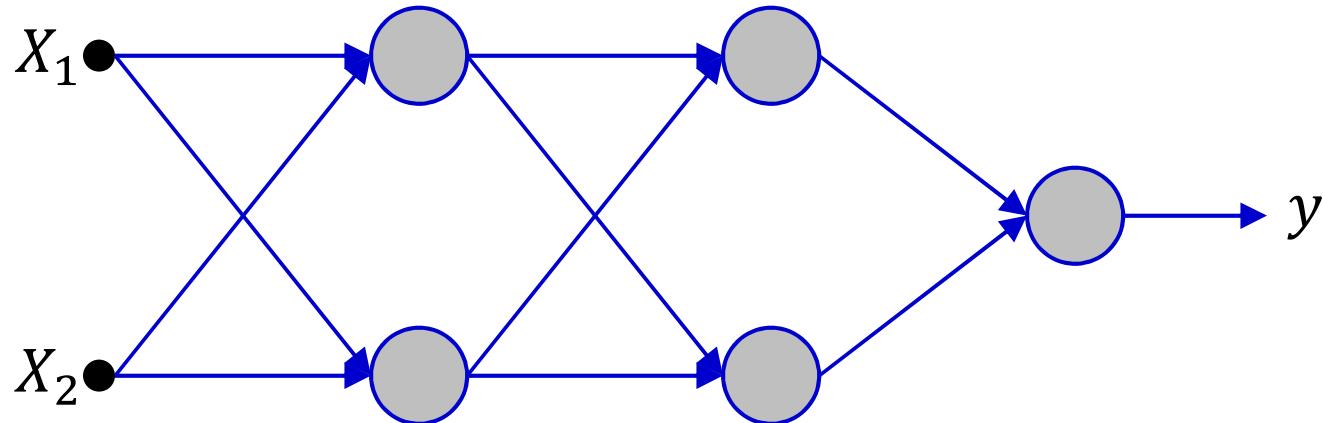


- Small perturbations in  $x$  cause small perturbations in each of  $g_1 \dots g_M$ , each of which individually additively perturbs  $y$

# Returning to our problem

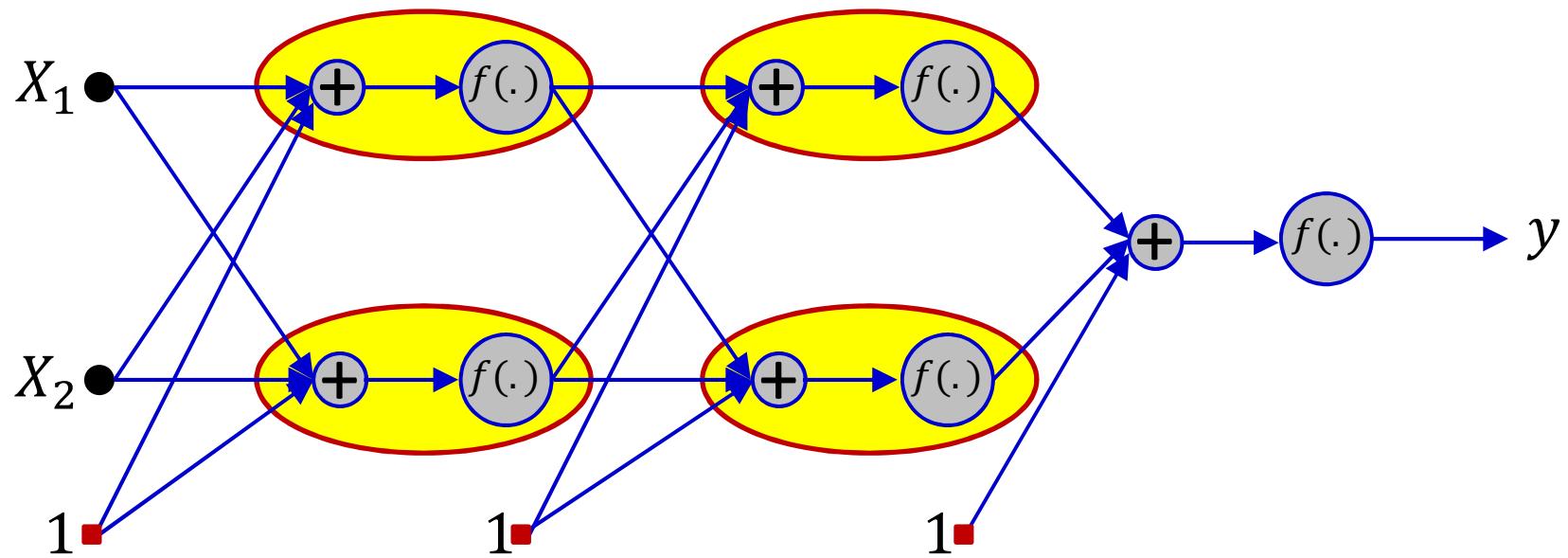
- How to compute  $\frac{d\text{Div}(\mathbf{Y}, \mathbf{d})}{dw_{i,j}^{(k)}}$

# A first closer look at the network



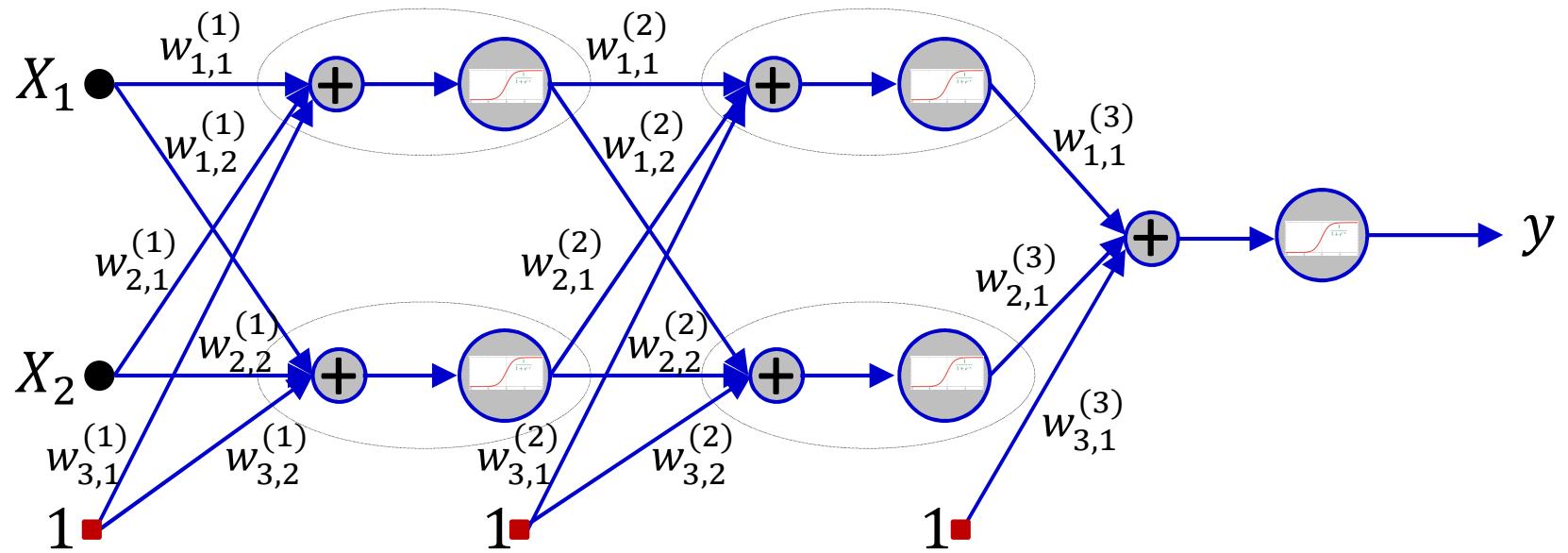
- Showing a tiny 2-input network for illustration
  - Actual network would have many more neurons and inputs

# A first closer look at the network



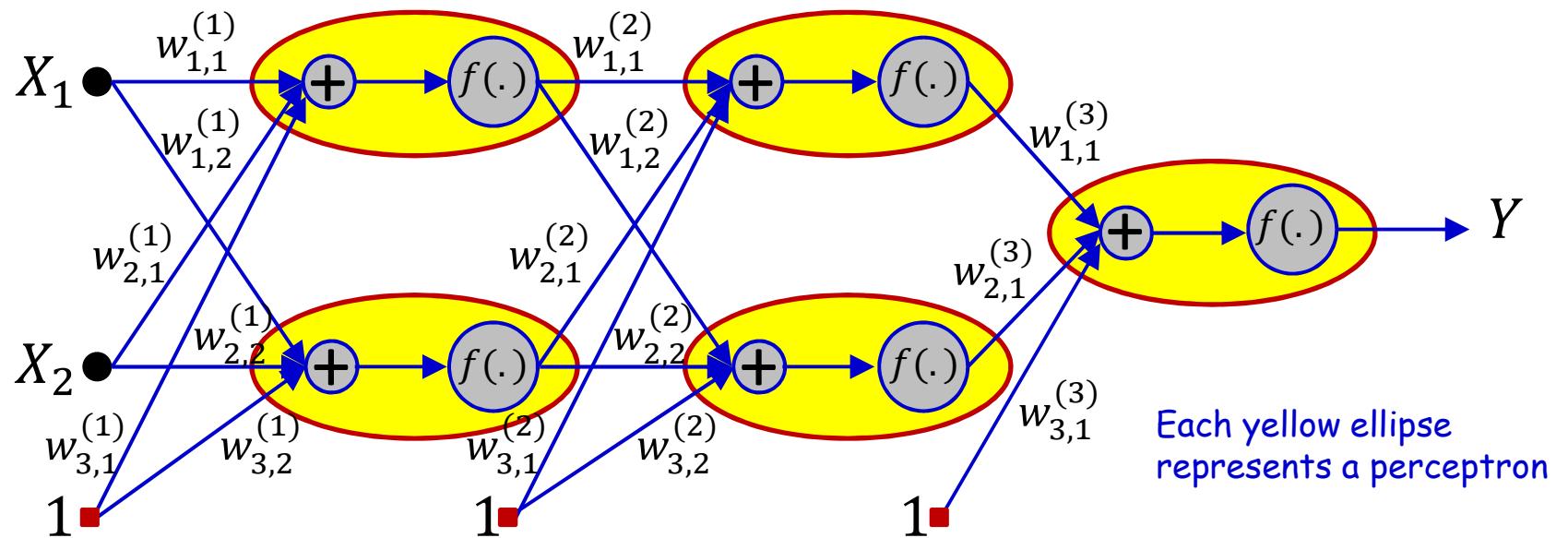
- Showing a tiny 2-input network for illustration
  - Actual network would have many more neurons and inputs
- Explicitly separating the weighted sum of inputs from the activation

# A first closer look at the network



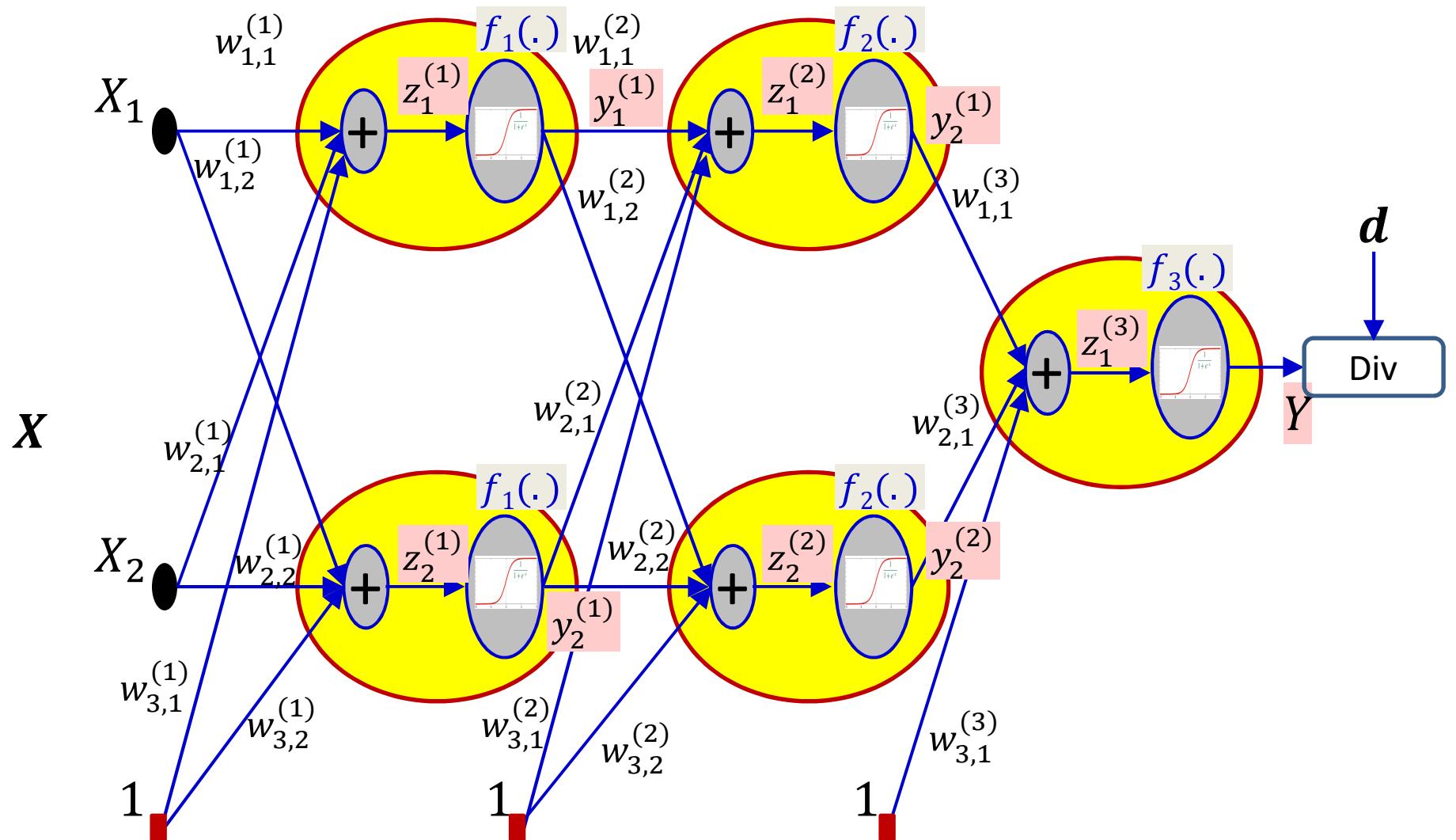
- Showing a tiny 2-input network for illustration
  - Actual network would have many more neurons and inputs
- Expanded **with all weights and activations shown**
- The overall function is differentiable w.r.t every weight, bias and input

# Computing the derivative for a *single* input



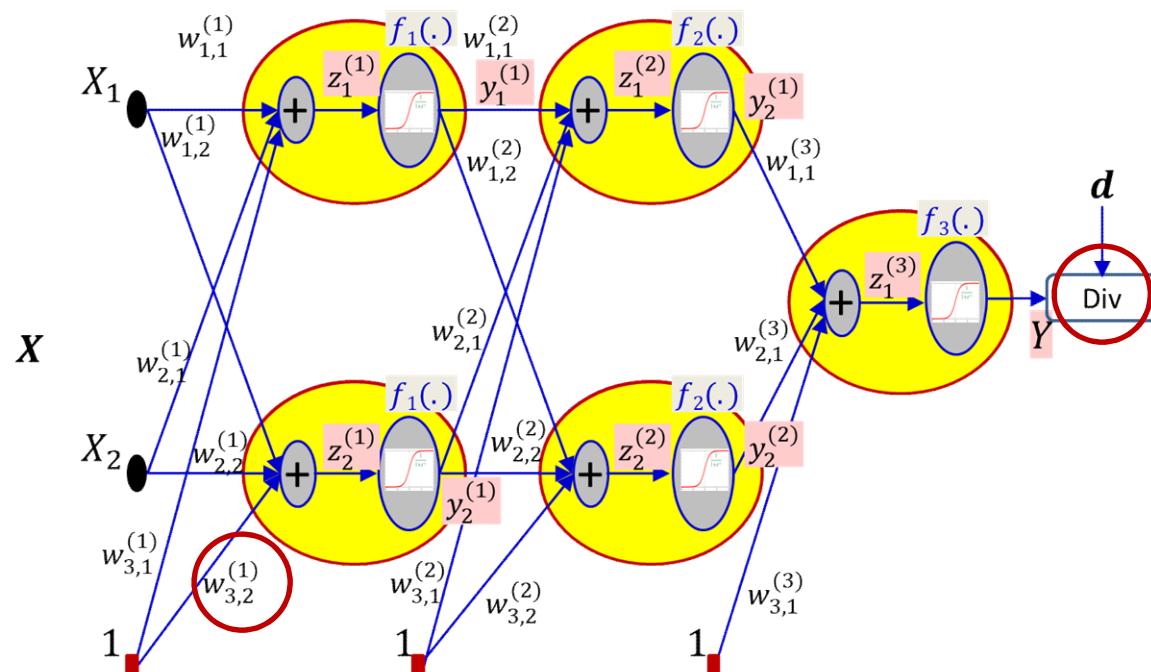
- Aim: compute derivative of  $\text{Div}(Y, d)$  w.r.t. each of the weights
- But first, lets label *all* our variables and activation functions

# Computing the derivative for a *single* input



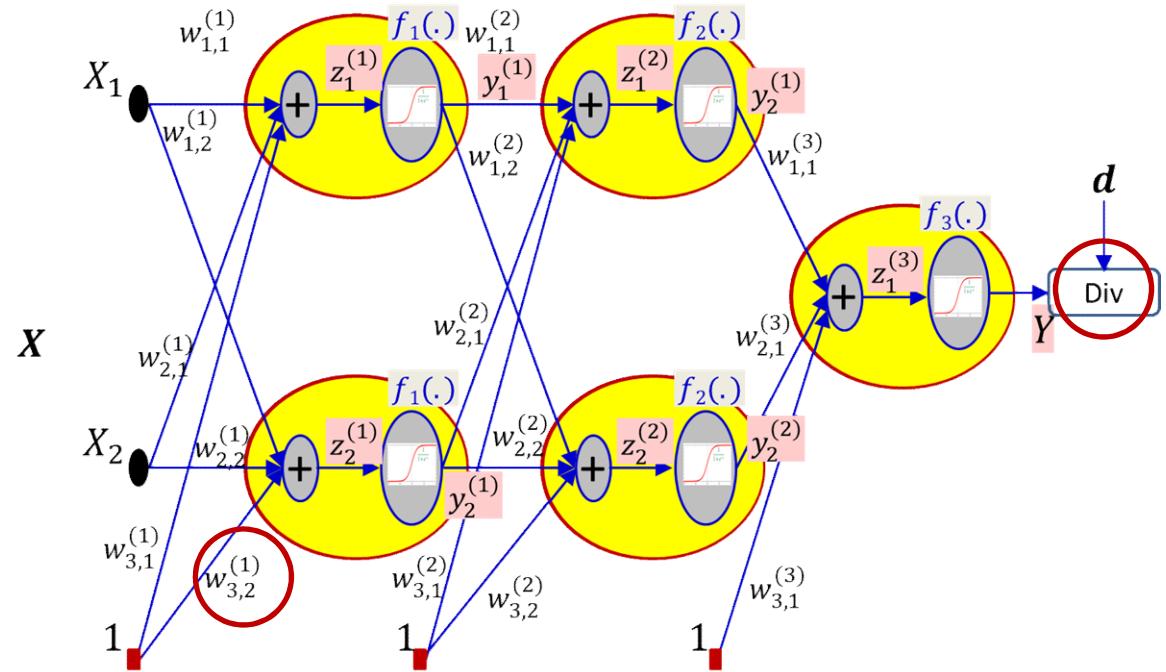
# Computing the gradient

- What is:  $\frac{d\text{Div}(Y, d)}{dw_{i,j}^{(k)}}$



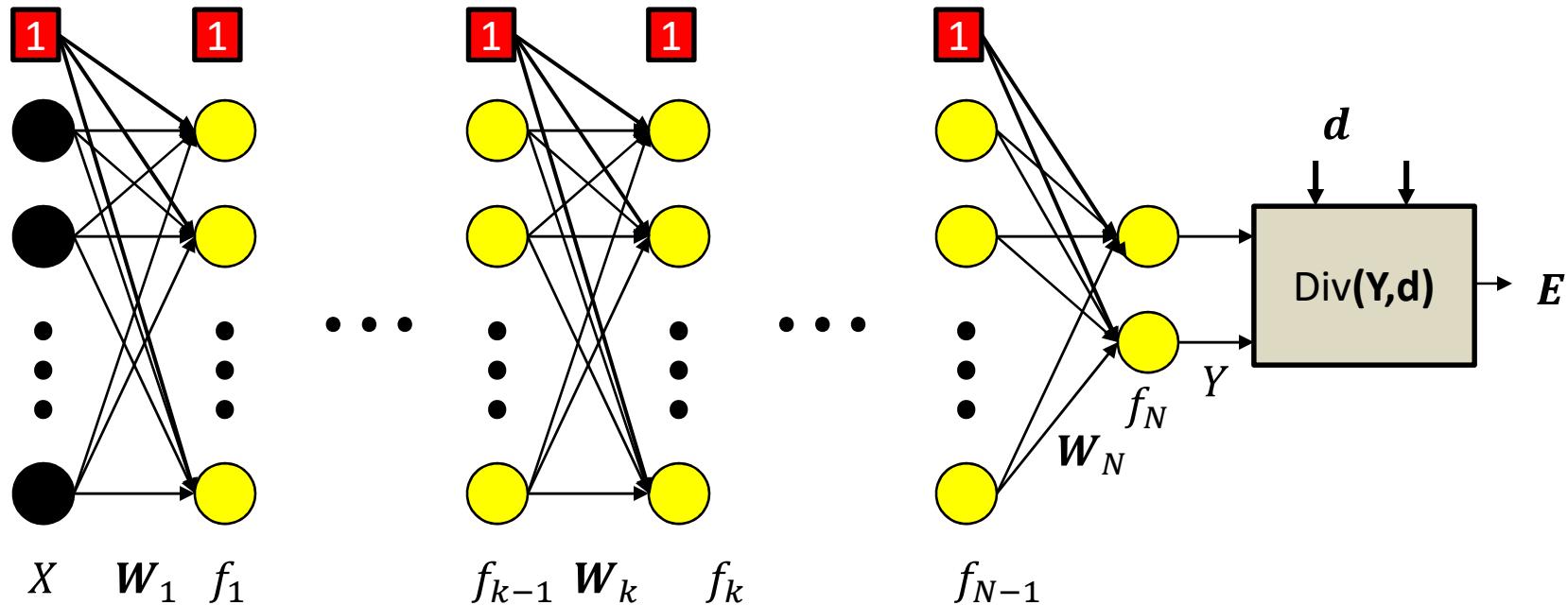
# Computing the gradient

- What is:  $\frac{d\text{Div}(Y, d)}{dw_{i,j}^{(k)}}$



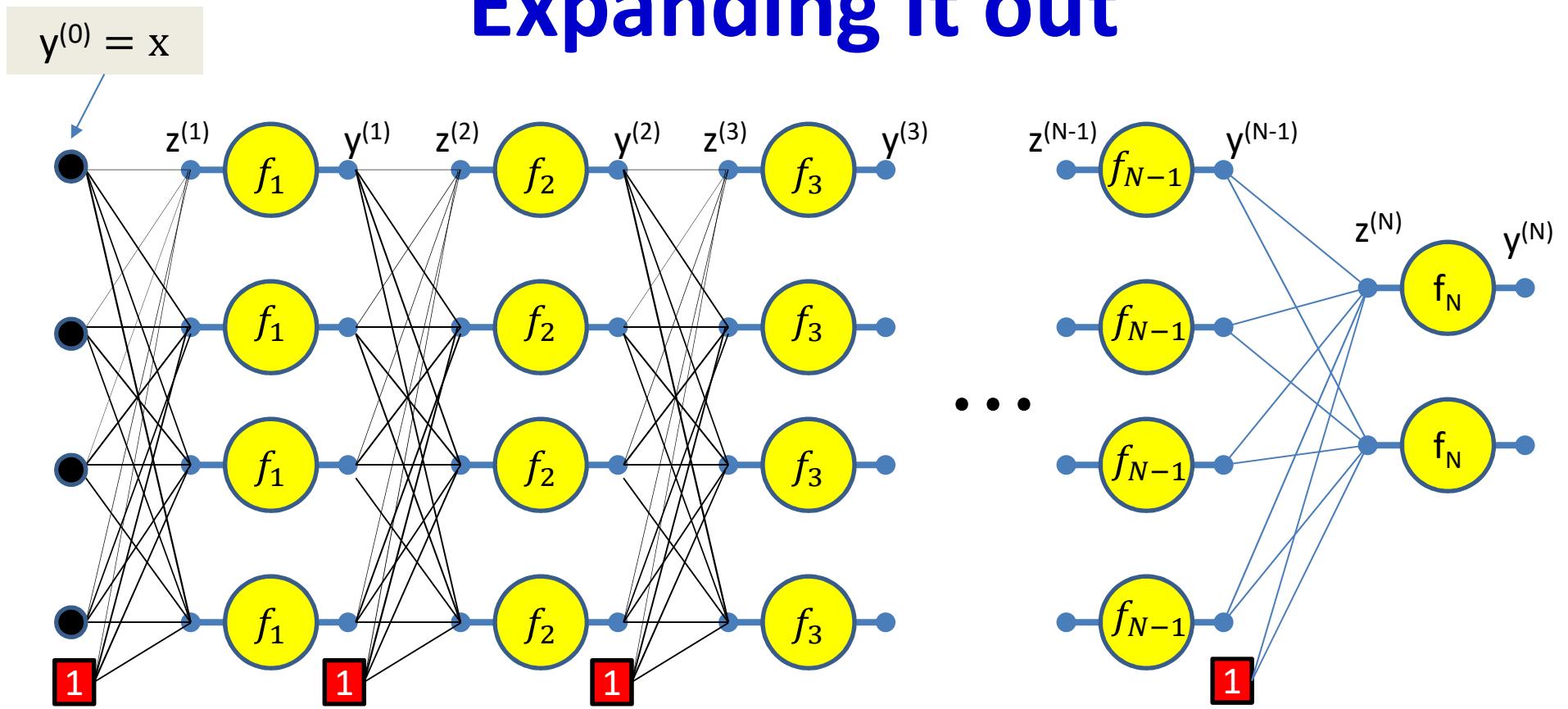
- Note: computation of the derivative requires intermediate and final output values of the network in response to the input

# BP: Scalar Formulation



- The network again

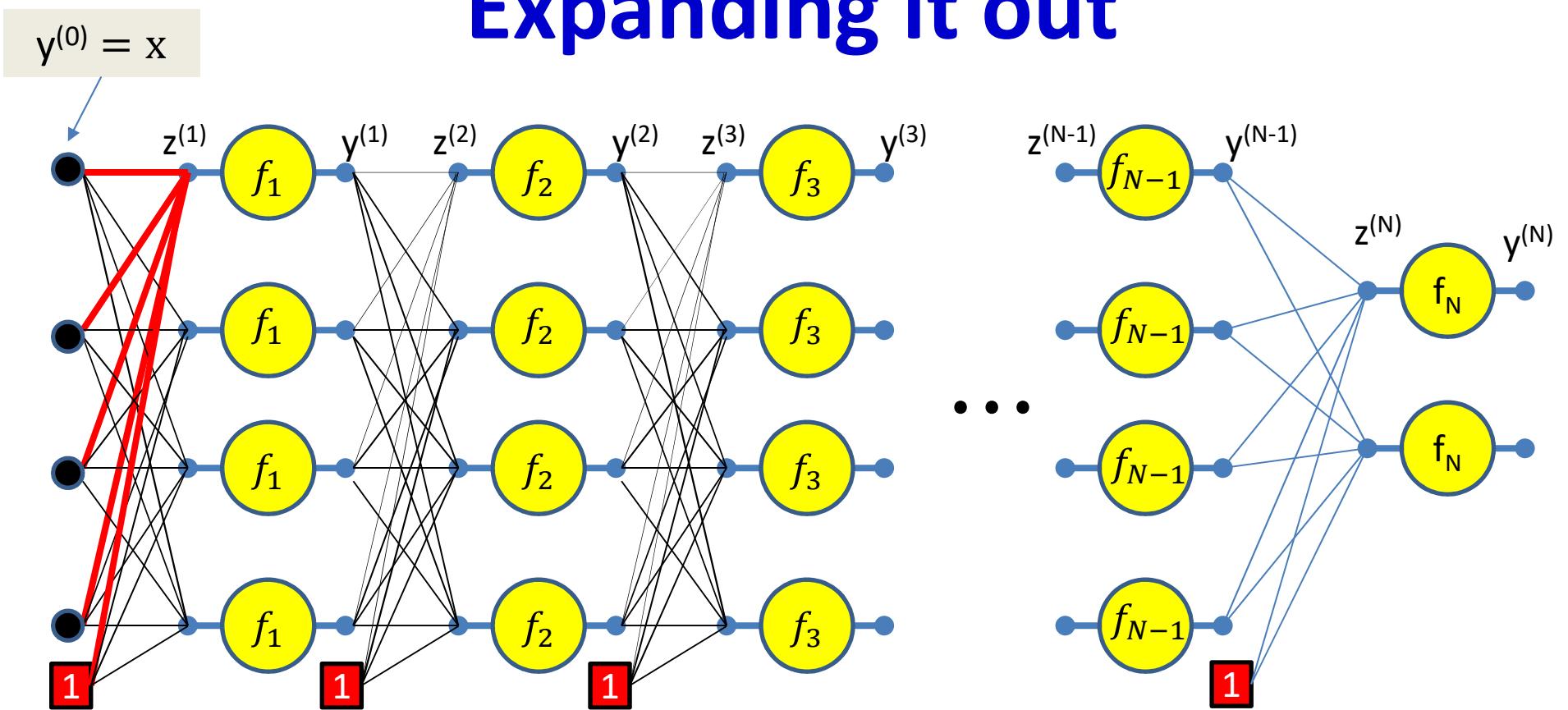
# Expanding it out



Setting  $y_i^{(0)} = x_i$  for notational convenience

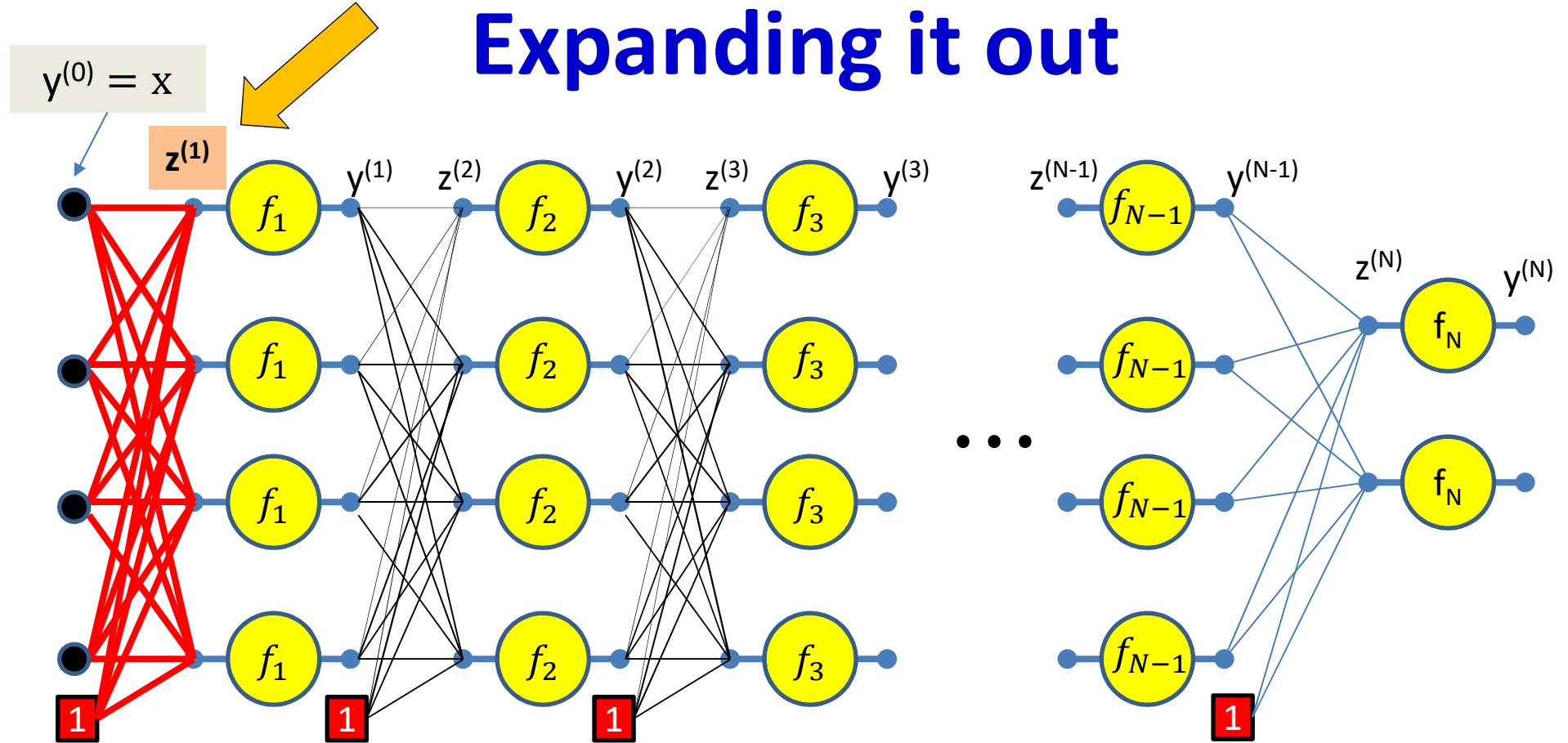
Assuming  $w_{0j}^{(k)} = b_j^{(k)}$  and  $y_0^{(k)} = 1$  -- assuming the bias is a weight and extending the output of every layer by a constant 1, to account for the biases

# Expanding it out

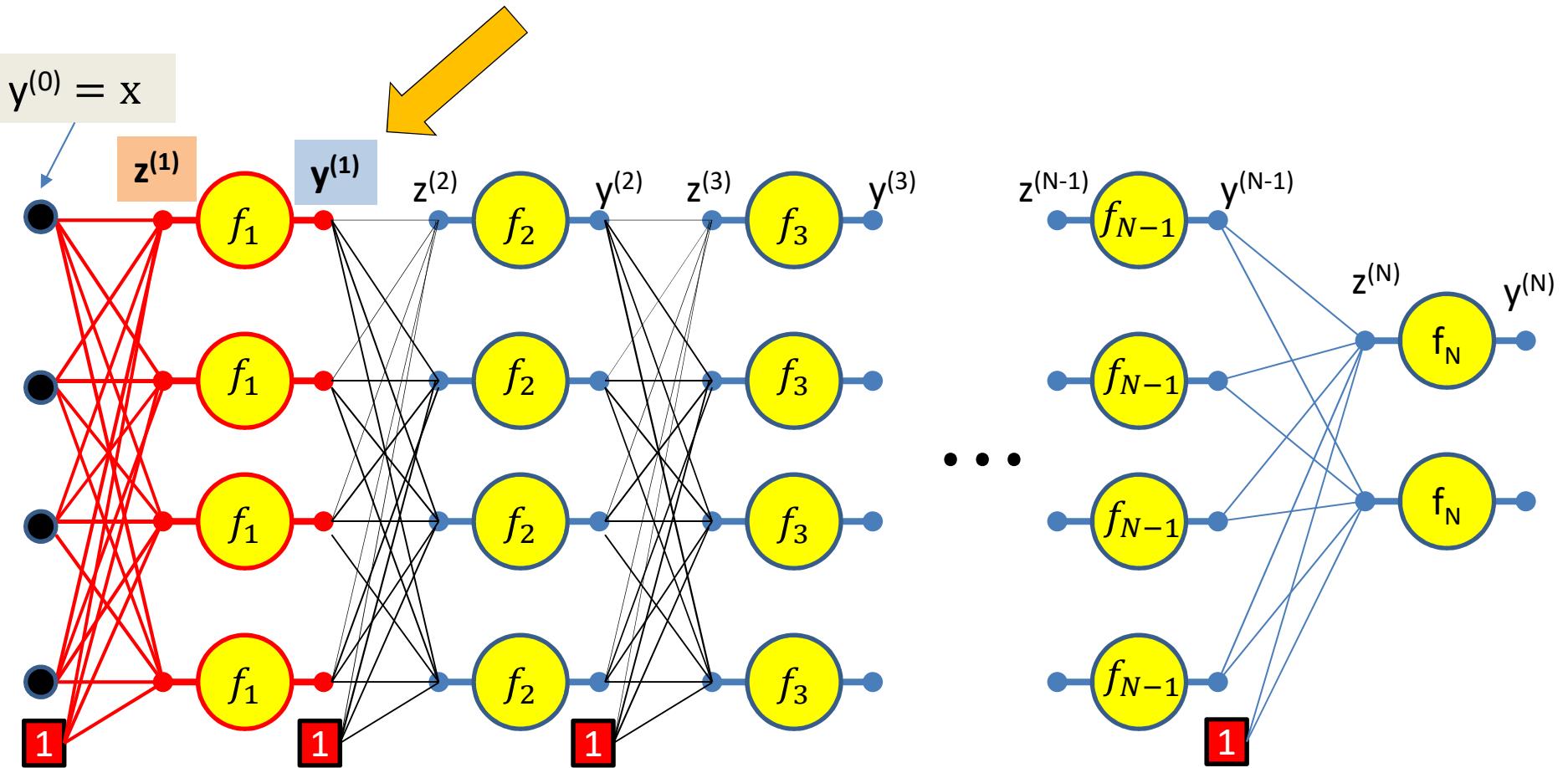


$$z_1^{(1)} = \sum_i w_{i1}^{(1)} y_i^{(0)}$$

# Expanding it out

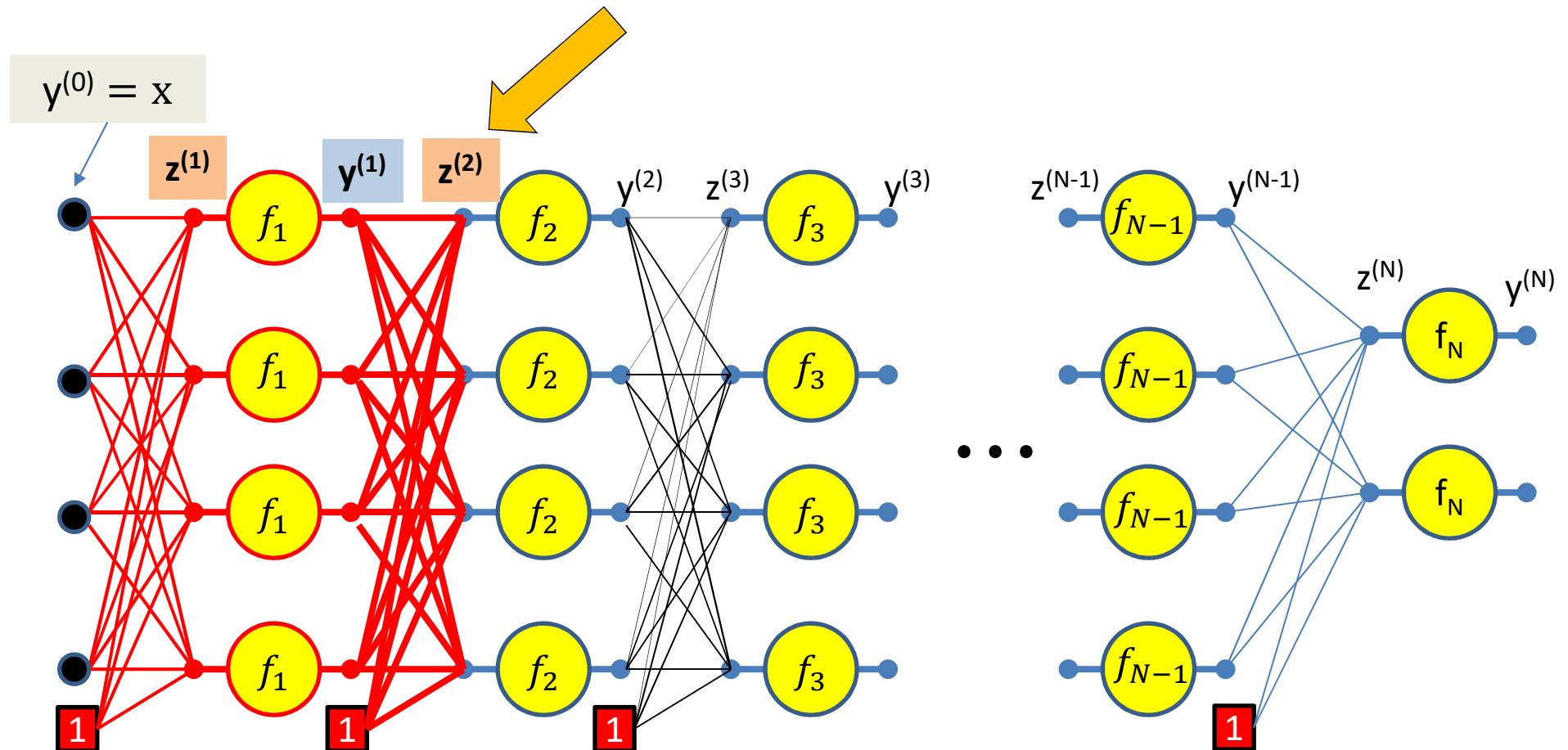


$$z_j^{(1)} = \sum_i w_{ij}^{(1)} y_i^{(0)}$$



$$z_j^{(1)} = \sum_i w_{ij}^{(1)} y_i^{(0)}$$

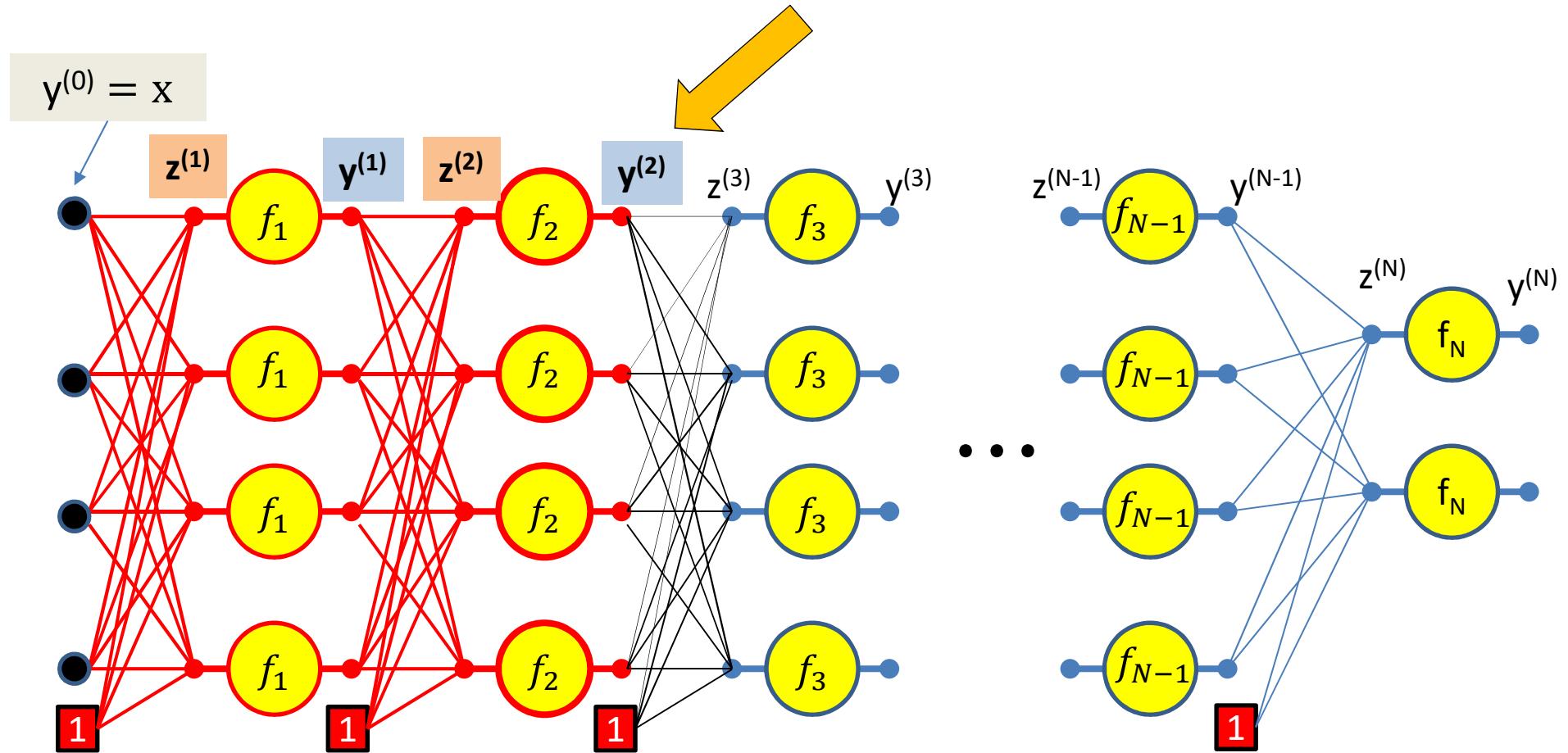
$$y_j^{(1)} = f_1(z_j^{(1)})$$



$$z_j^{(1)} = \sum_i w_{ij}^{(1)} y_i^{(0)}$$

$$y_j^{(1)} = f_1(z_j^{(1)})$$

$$z_j^{(2)} = \sum_i w_{ij}^{(2)} y_i^{(1)}$$

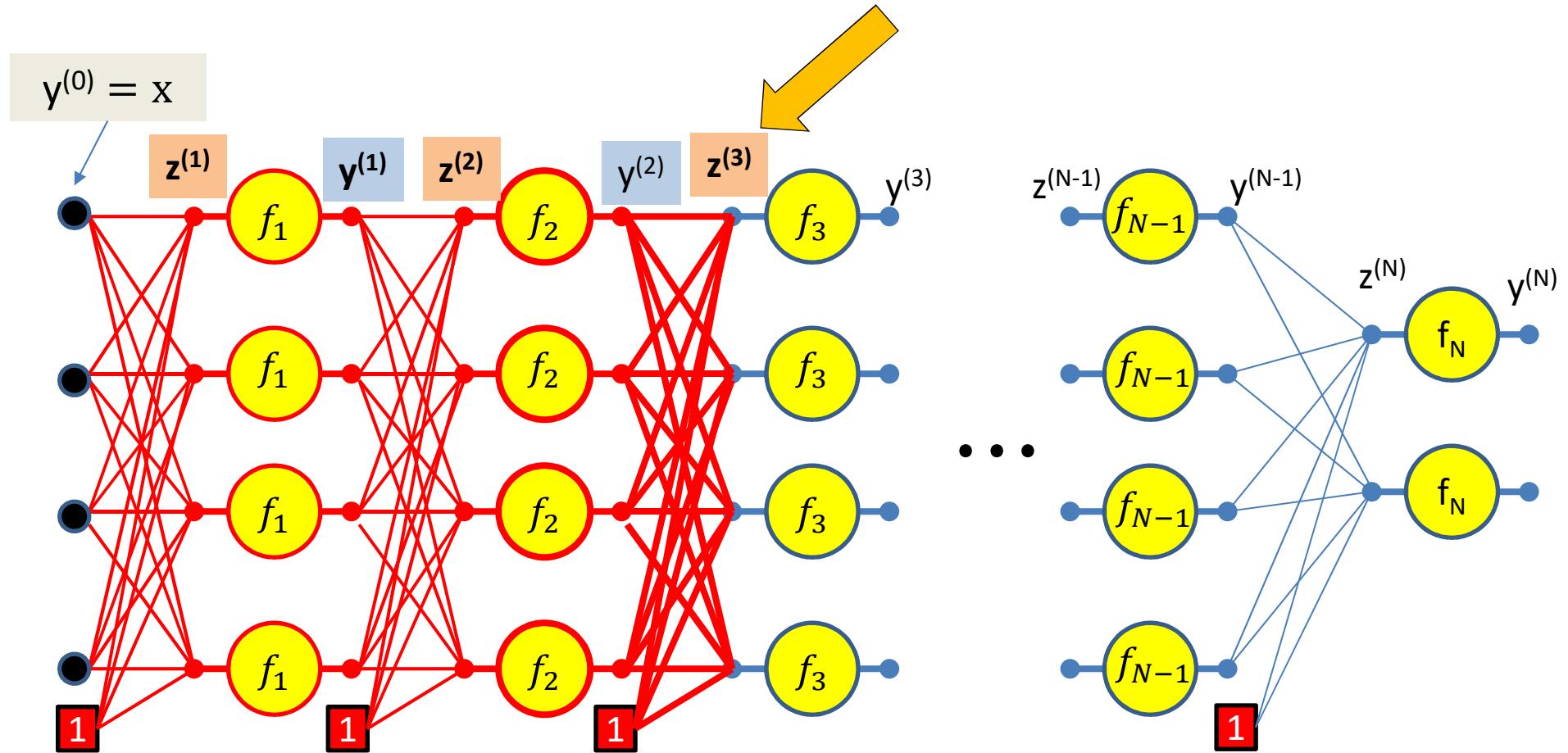


$$z_j^{(1)} = \sum_i w_{ij}^{(1)} y_i^{(0)}$$

$$y_j^{(1)} = f_1(z_j^{(1)})$$

$$z_j^{(2)} = \sum_i w_{ij}^{(2)} y_i^{(1)}$$

$$y_j^{(2)} = f_2(z_j^{(2)})$$



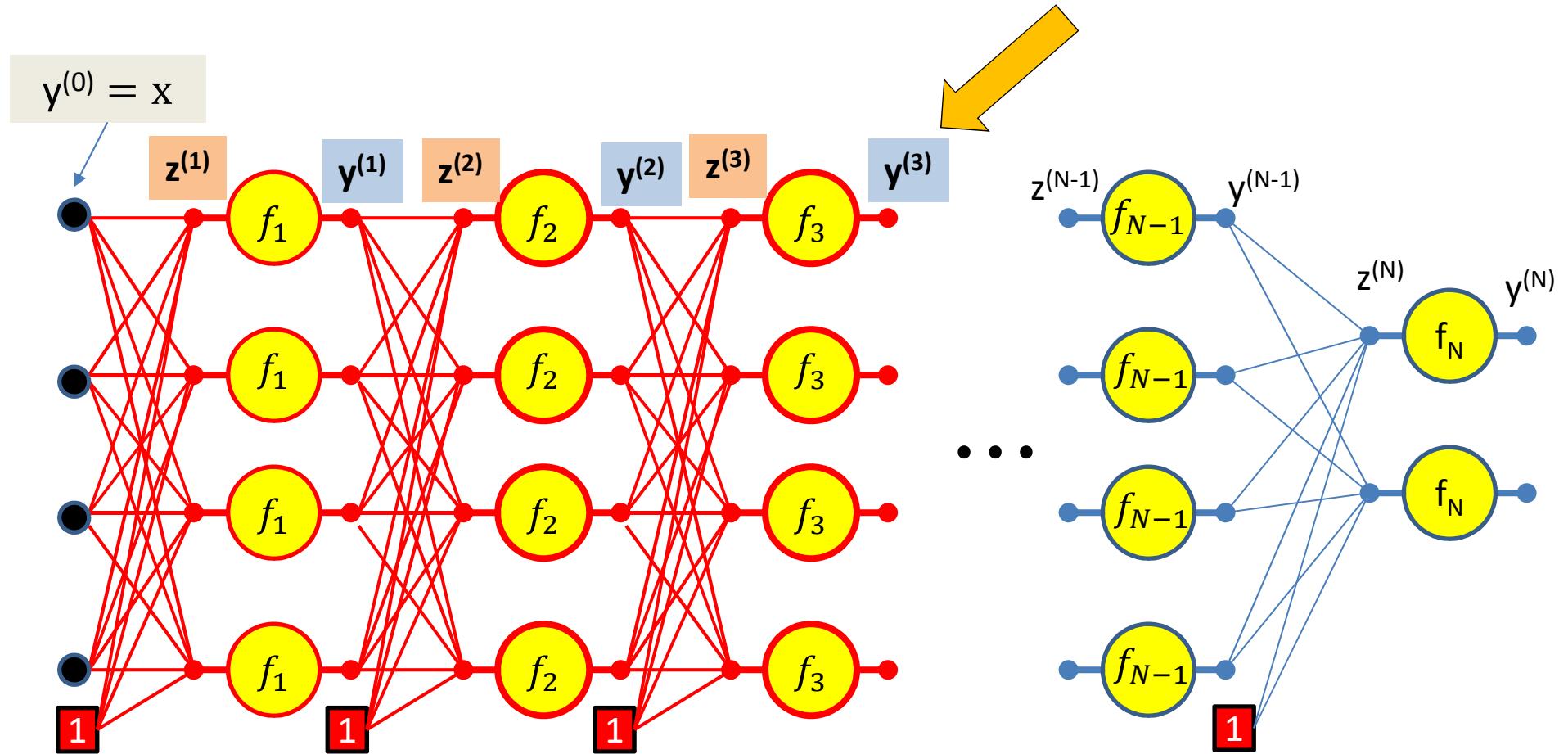
$$z_j^{(1)} = \sum_i w_{ij}^{(1)} y_i^{(0)}$$

$$y_j^{(1)} = f_1(z_j^{(1)})$$

$$z_j^{(2)} = \sum_i w_{ij}^{(2)} y_i^{(1)}$$

$$y_j^{(2)} = f_2(z_j^{(2)})$$

$$z_j^{(3)} = \sum_i w_{ij}^{(3)} y_i^{(2)}$$



$$z_j^{(1)} = \sum_i w_{ij}^{(1)} y_i^{(0)}$$

$$y_j^{(1)} = f_1(z_j^{(1)})$$

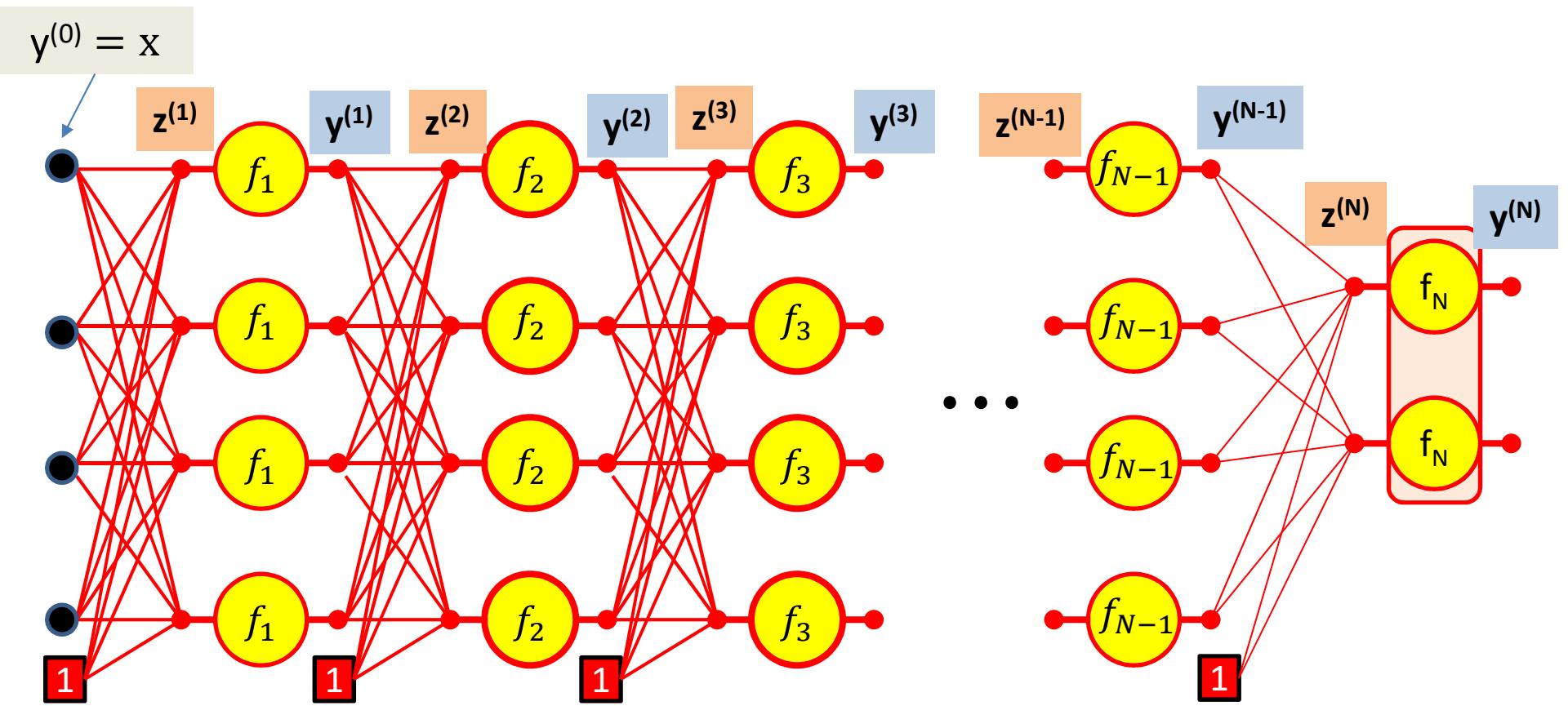
$$z_j^{(2)} = \sum_i w_{ij}^{(2)} y_i^{(1)}$$

$$y_j^{(2)} = f_2(z_j^{(2)})$$

$$z_j^{(3)} = \sum_i w_{ij}^{(3)} y_i^{(2)}$$

$$y_j^{(3)} = f_3(z_j^{(3)})$$

$\dots$

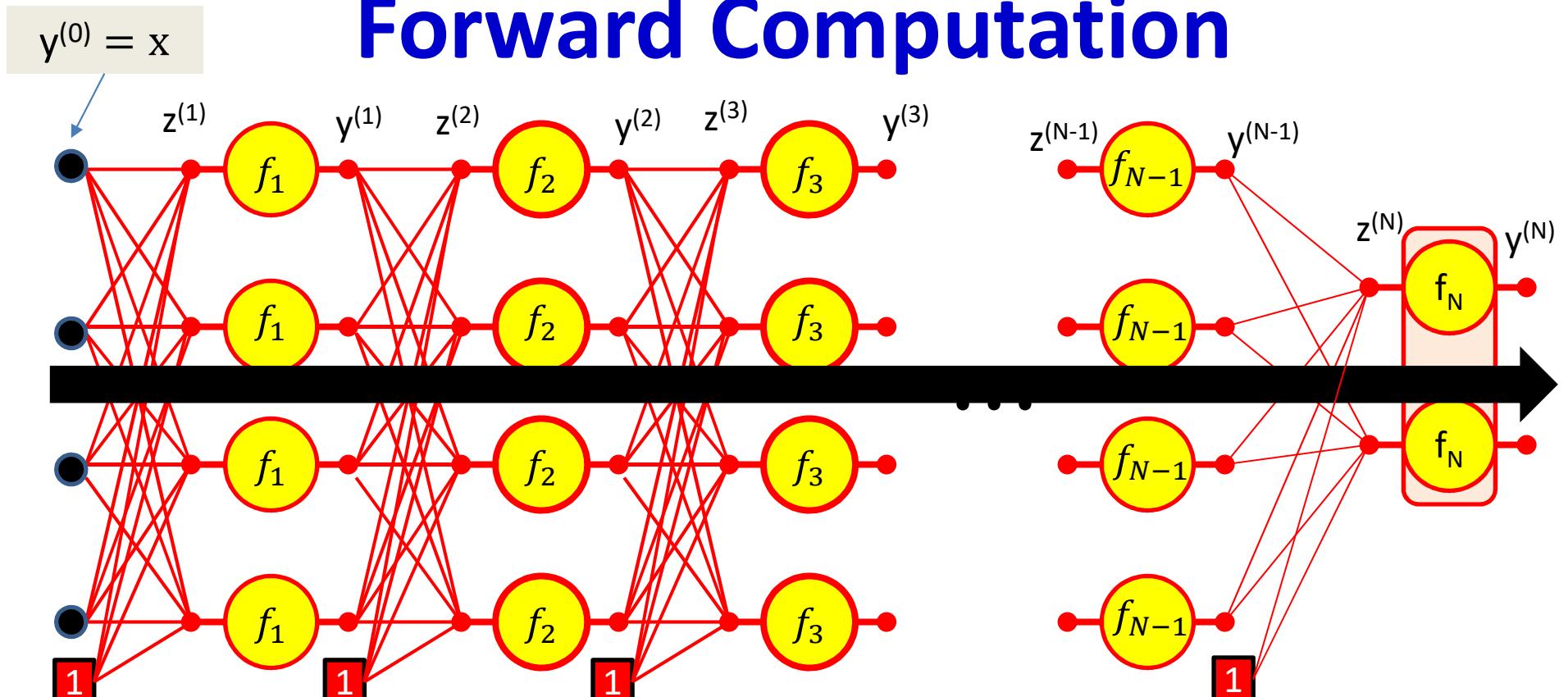


$$y_j^{(N-1)} = f_{N-1}(z_j^{(N-1)})$$

$$z_j^{(N)} = \sum_i w_{ij}^{(N)} y_i^{(N-1)}$$

$$\mathbf{y}^{(N)} = f_N(\mathbf{z}^{(N)})$$

# Forward Computation



ITERATE FOR  $k = 1:N$

for  $j = 1:\text{layer-width}$

$$y_i^{(0)} = x_i$$

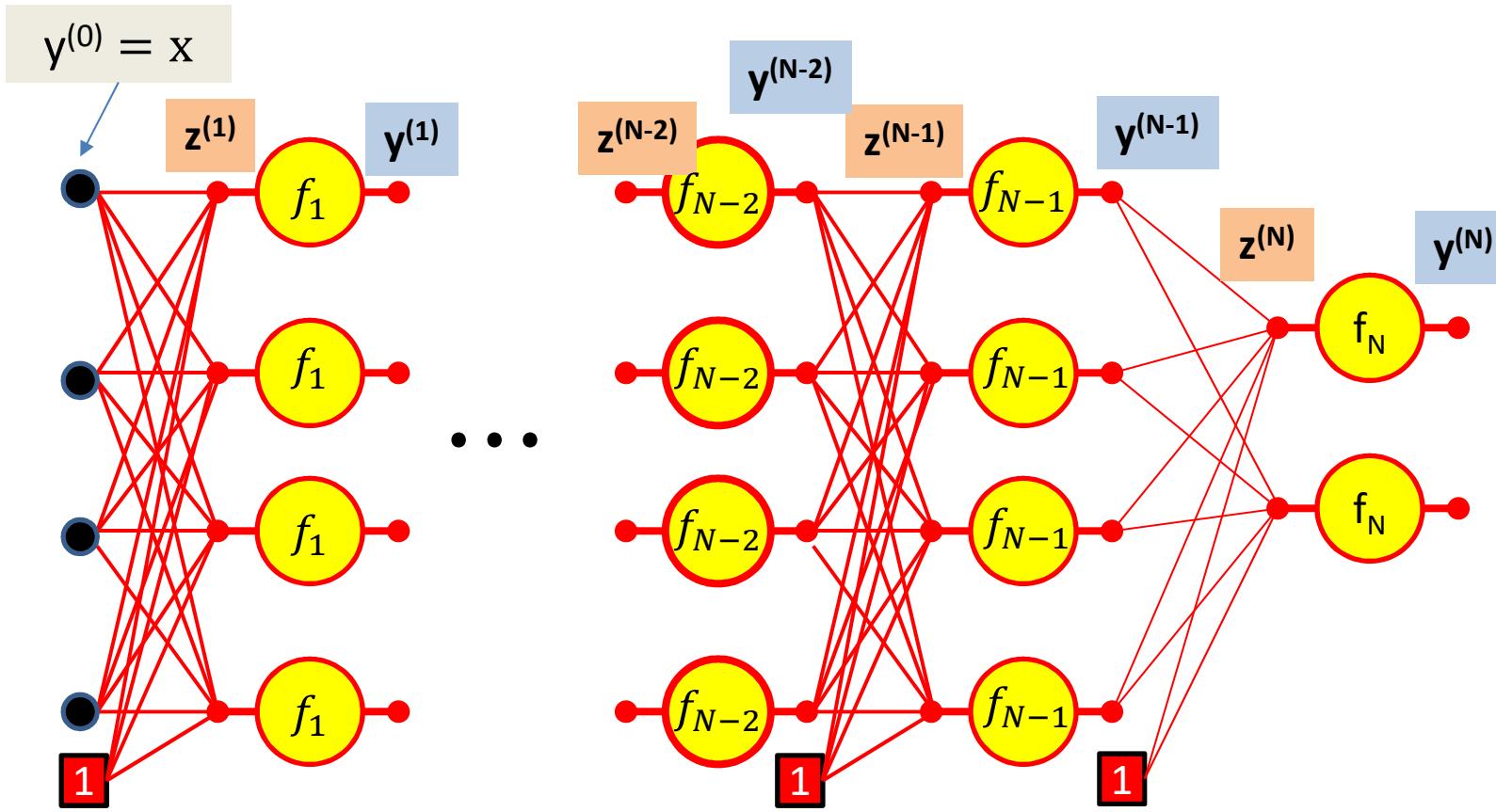
$$z_j^{(k)} = \sum_i w_{ij}^{(k)} y_i^{(k-1)}$$

$$y_j^{(k)} = f_k(z_j^{(k)})$$

# Forward “Pass”

- Input:  $D$  dimensional vector  $\mathbf{x} = [x_j, j = 1 \dots D]$
- Set:
  - $D_0 = D$ , is the width of the 0<sup>th</sup> (input) layer
  - $y_j^{(0)} = x_j, j = 1 \dots D; y_0^{(k=1\dots N)} = x_0 = 1$
- For layer  $k = 1 \dots N$ 
  - For  $j = 1 \dots D_k$   $D_k$  is the size of the kth layer
    - $z_j^{(k)} = \sum_{i=0}^{D_{k-1}} w_{i,j}^{(k)} y_i^{(k-1)}$
    - $y_j^{(k)} = f_k(z_j^{(k)})$
- Output:
  - $Y = y_j^{(N)}, j = 1..D_N$

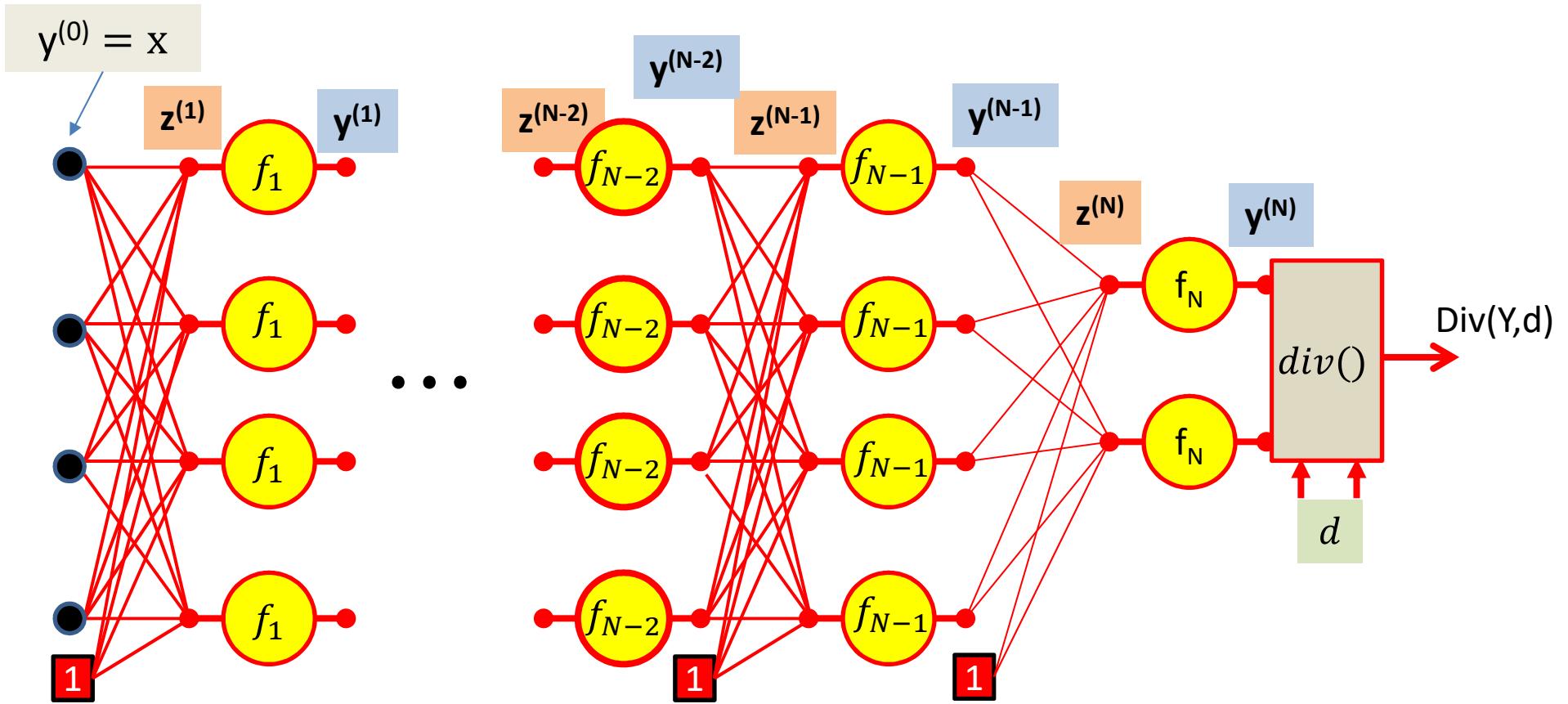
# Computing derivatives



We have computed all these intermediate values in the forward computation

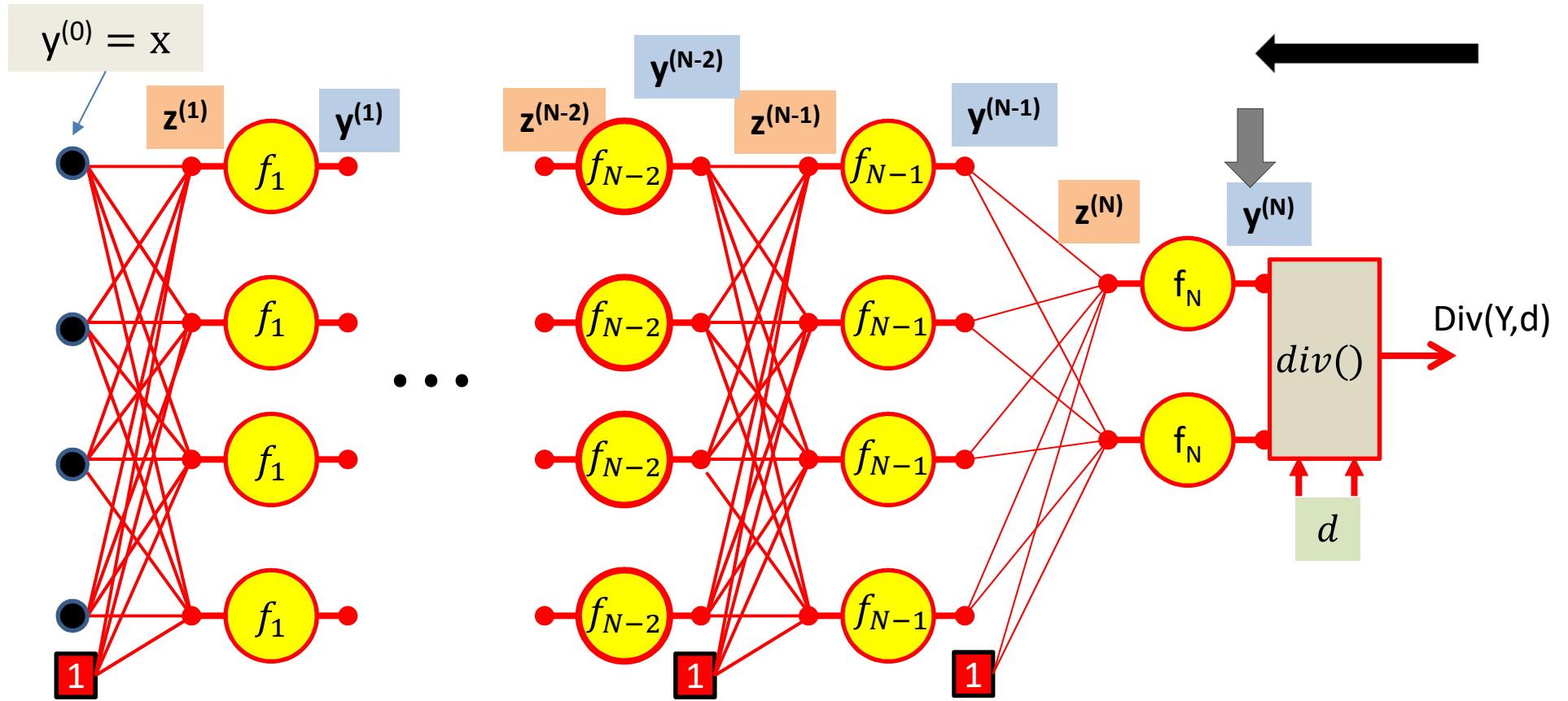
We must remember them - we will need them to compute the derivatives

# Computing derivatives



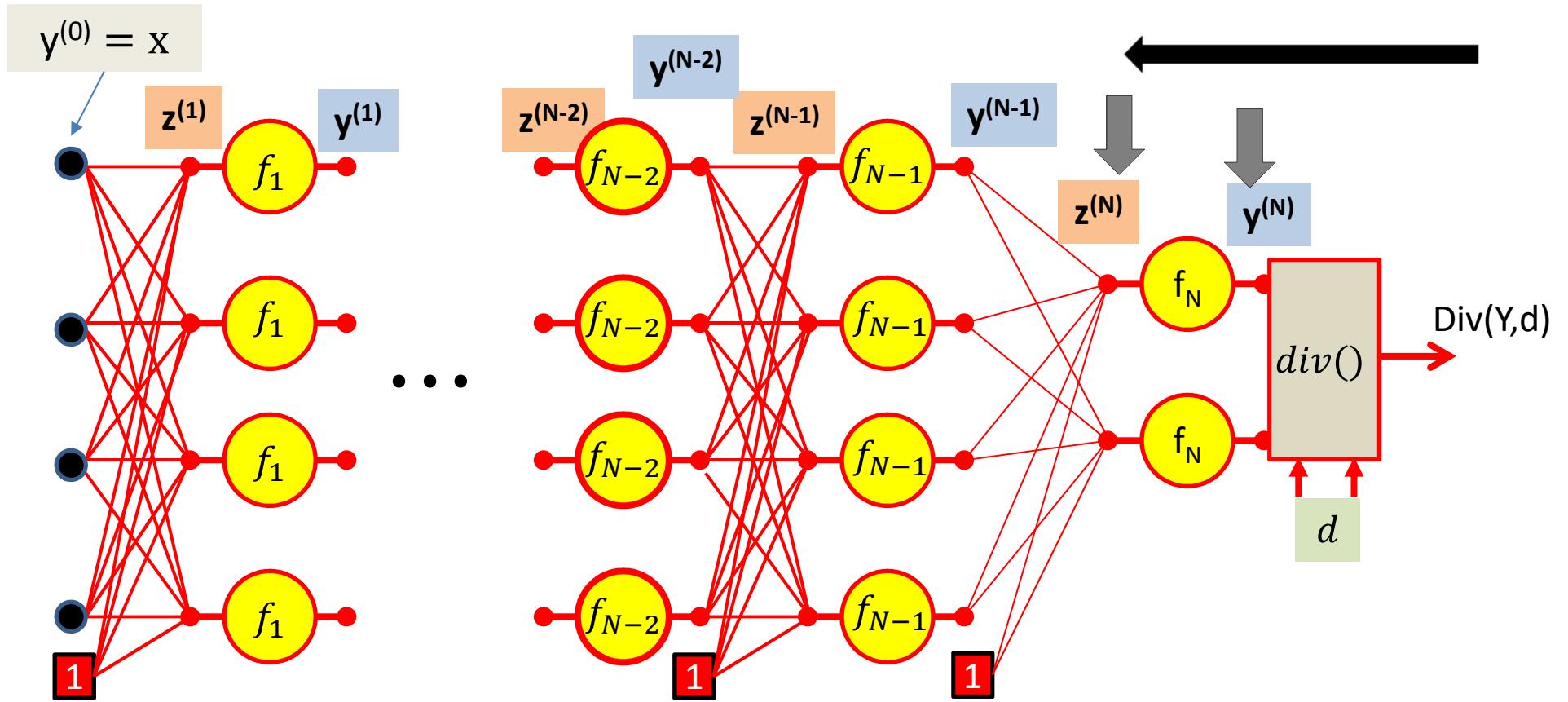
First, we compute the divergence between the output of the net  $y = y^{(N)}$  and the desired output  $d$

# Computing derivatives



We then compute  $\nabla_{Y^{(N)}} \text{div}(\cdot)$  the derivative of the divergence w.r.t. the final output of the network  $y^{(N)}$

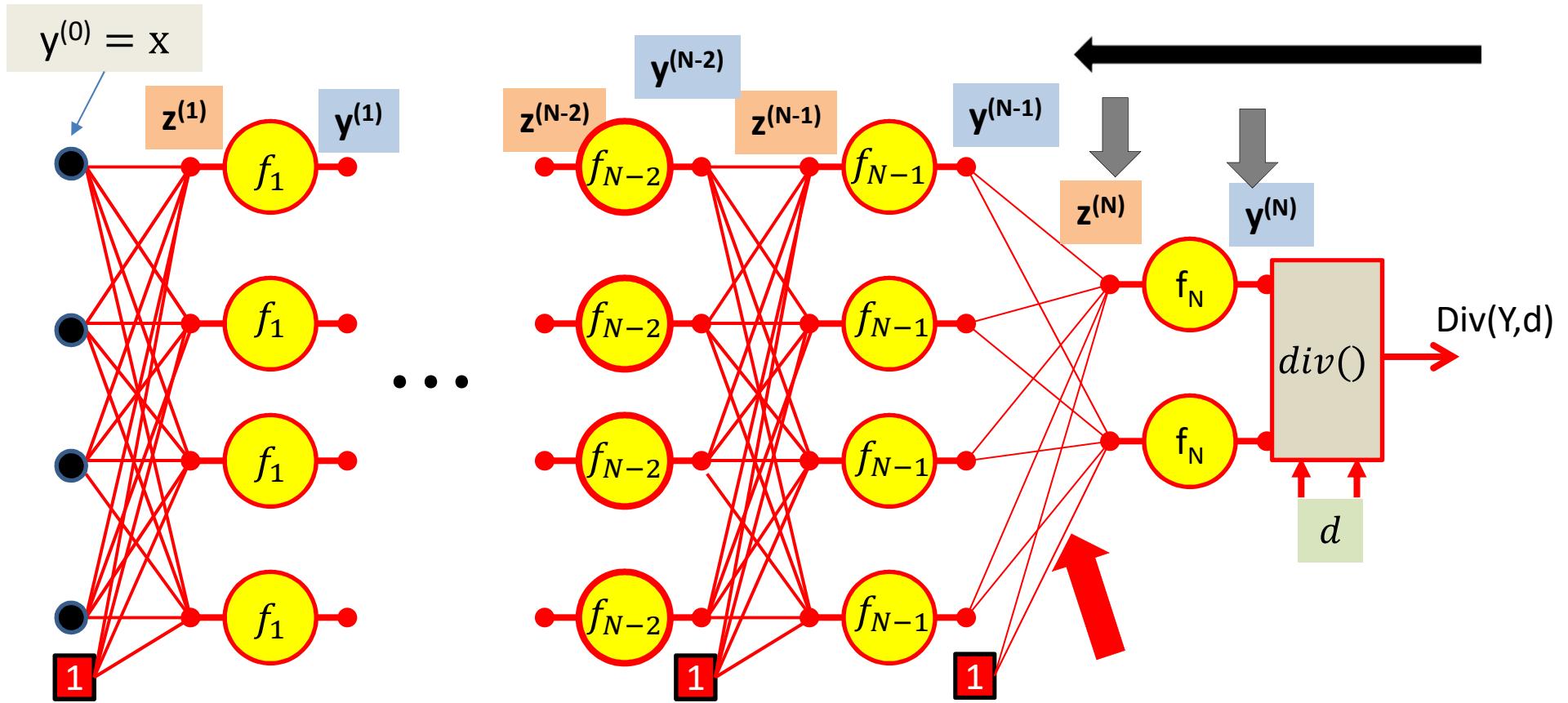
# Computing derivatives



We then compute  $\nabla_{Y^{(N)}} \text{div}(\cdot)$  the derivative of the divergence w.r.t. the final output of the network  $y^{(N)}$

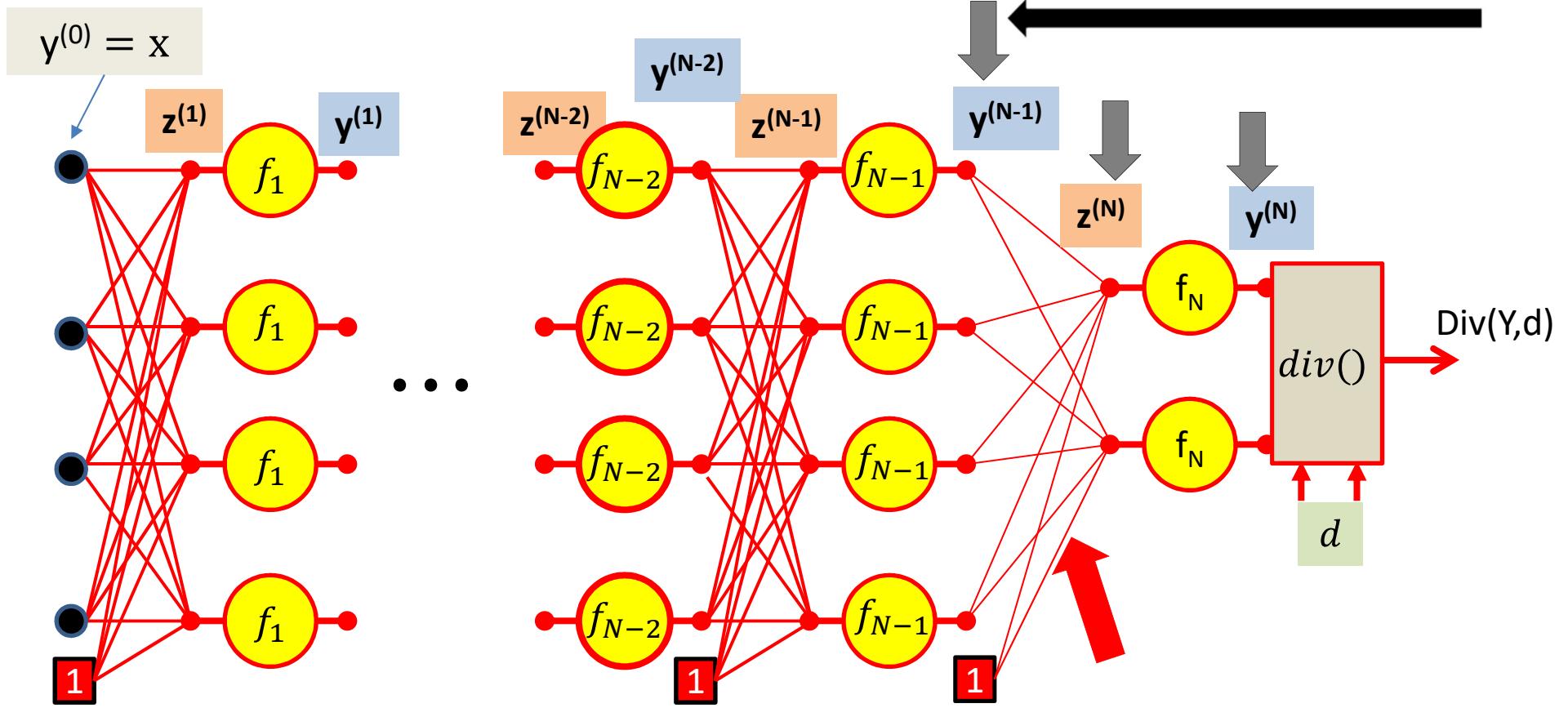
We then compute  $\nabla_{z^{(N)}} \text{div}(\cdot)$  the derivative of the divergence w.r.t. the *pre-activation affine combination*  $z^{(N)}$  using the chain rule

# Computing derivatives



Continuing on, we will compute  $\nabla_{W^{(N)}} \text{div}(\cdot)$  the derivative of the divergence with respect to the weights of the connections to the output layer

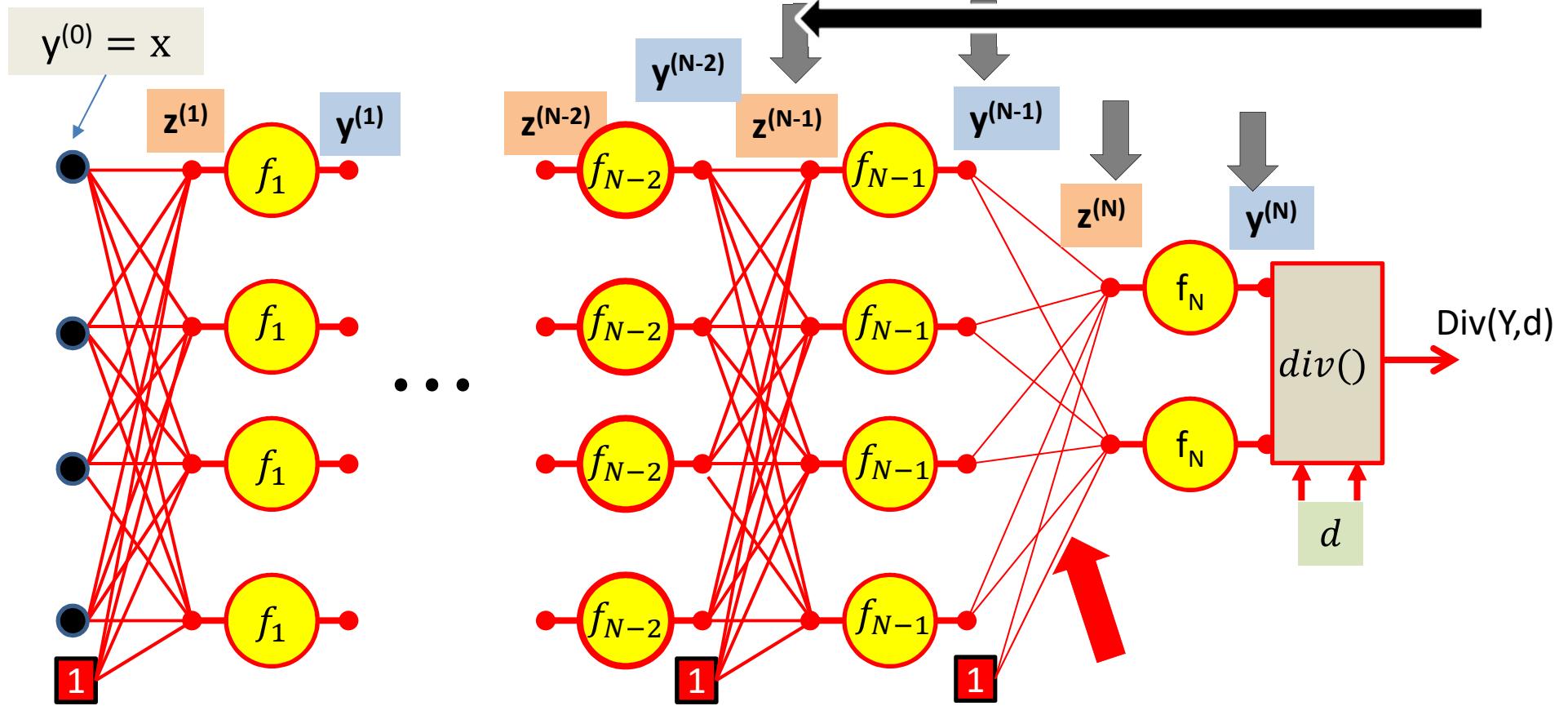
# Computing derivatives



Continuing on, we will compute  $\nabla_{W^{(N)}} \text{div}(\cdot)$  the derivative of the divergence with respect to the weights of the connections to the output layer

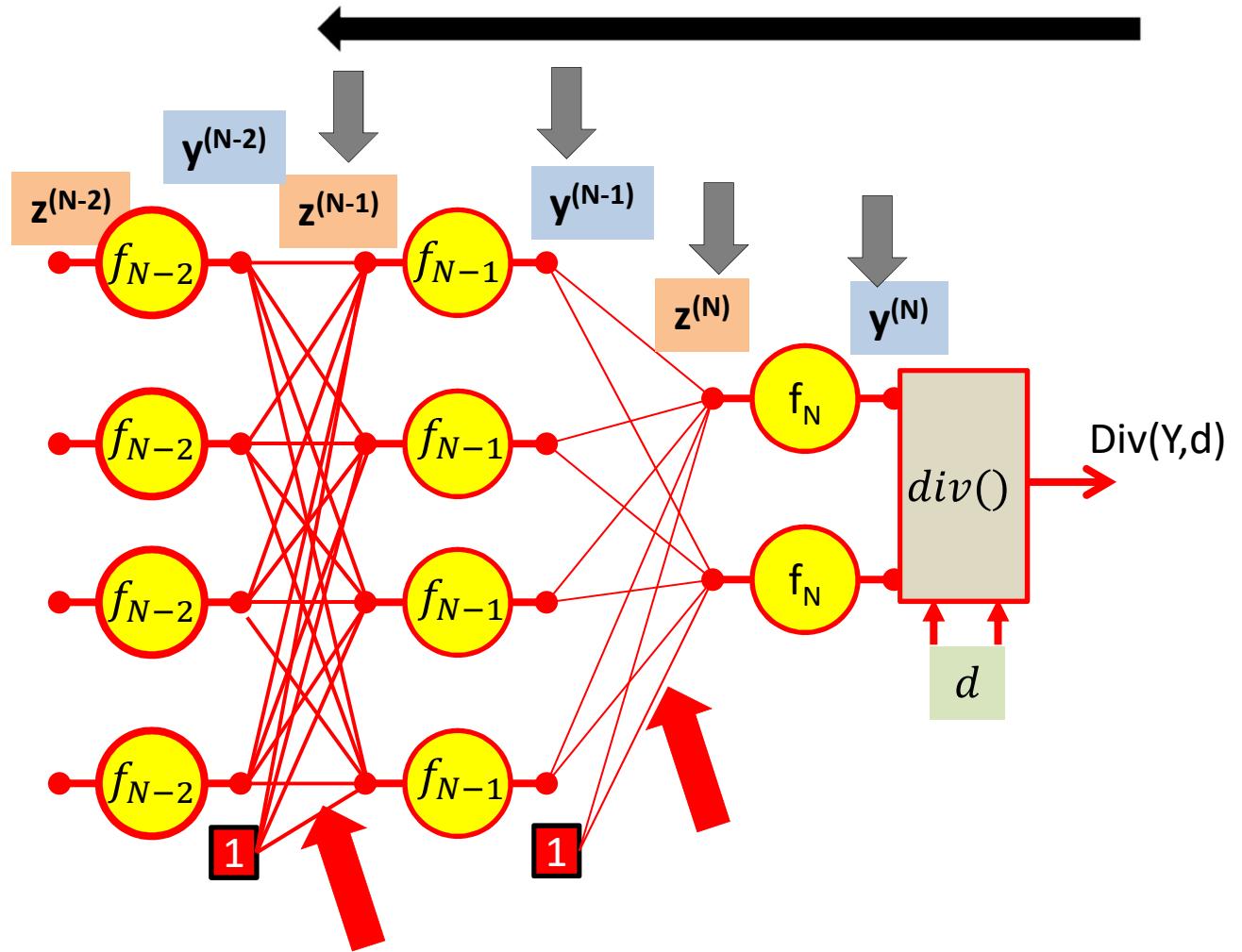
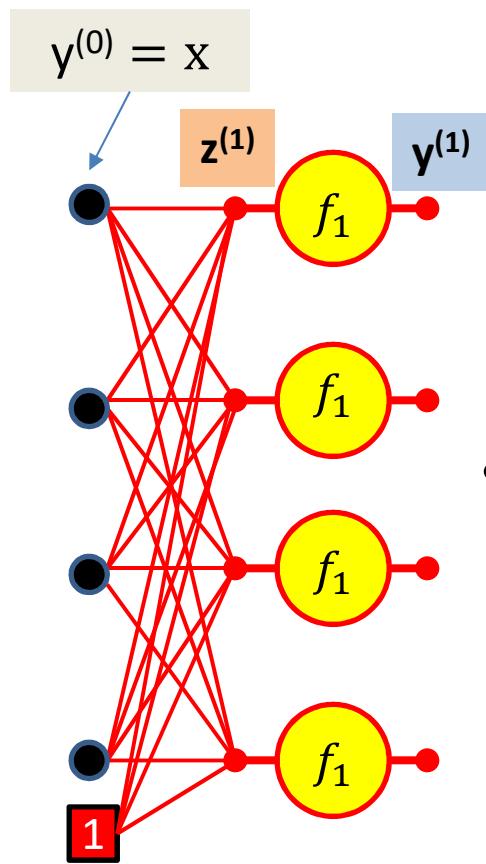
Then continue with the chain rule to compute  $\nabla_{Y^{(N-1)}} \text{div}(\cdot)$  the derivative of the divergence w.r.t. the output of the N-1th layer

# Computing derivatives



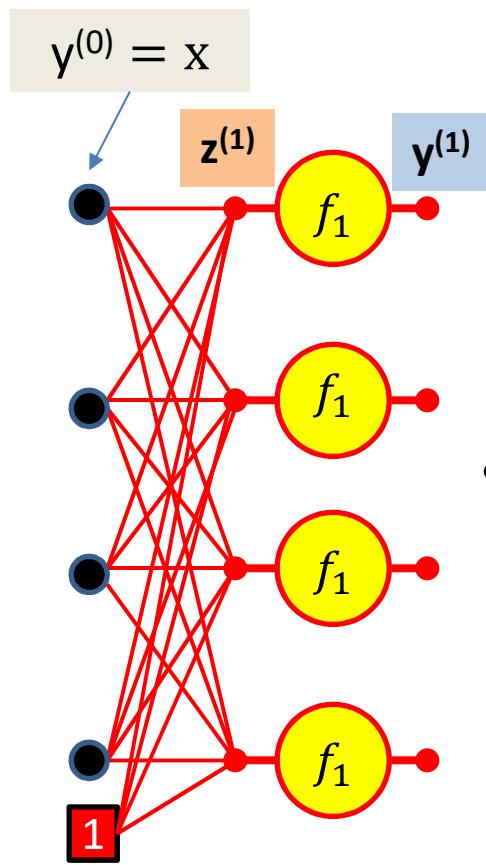
We continue our way backwards in the order shown

$$\nabla_{z^{(N-1)}} div(.)$$

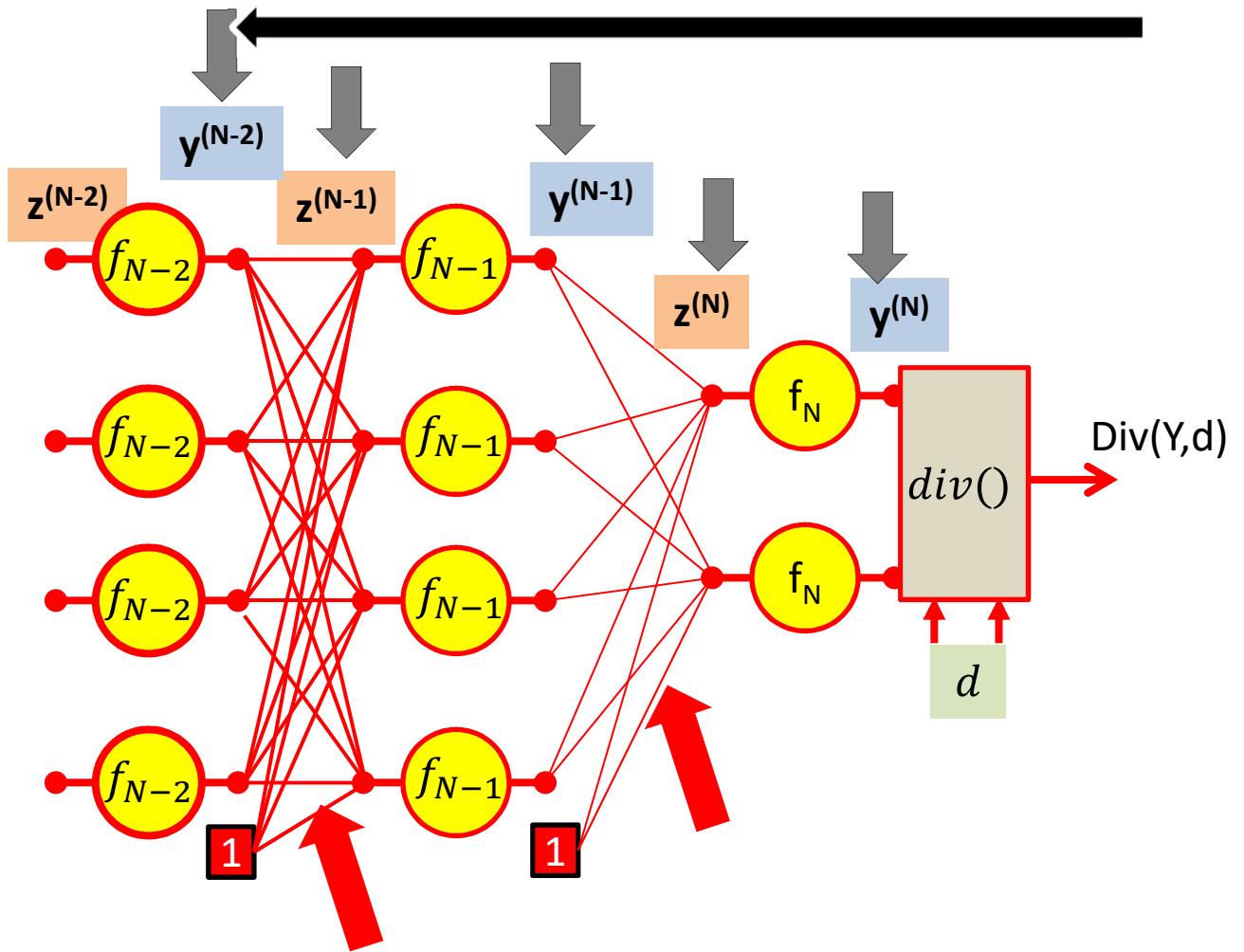


We continue our way backwards in the order shown

$$\nabla_{W^{(N-1)}} div(.)$$

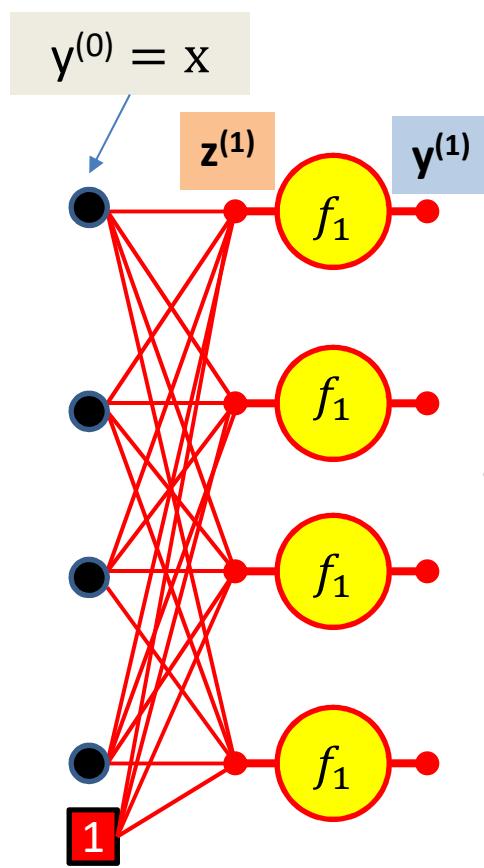


...

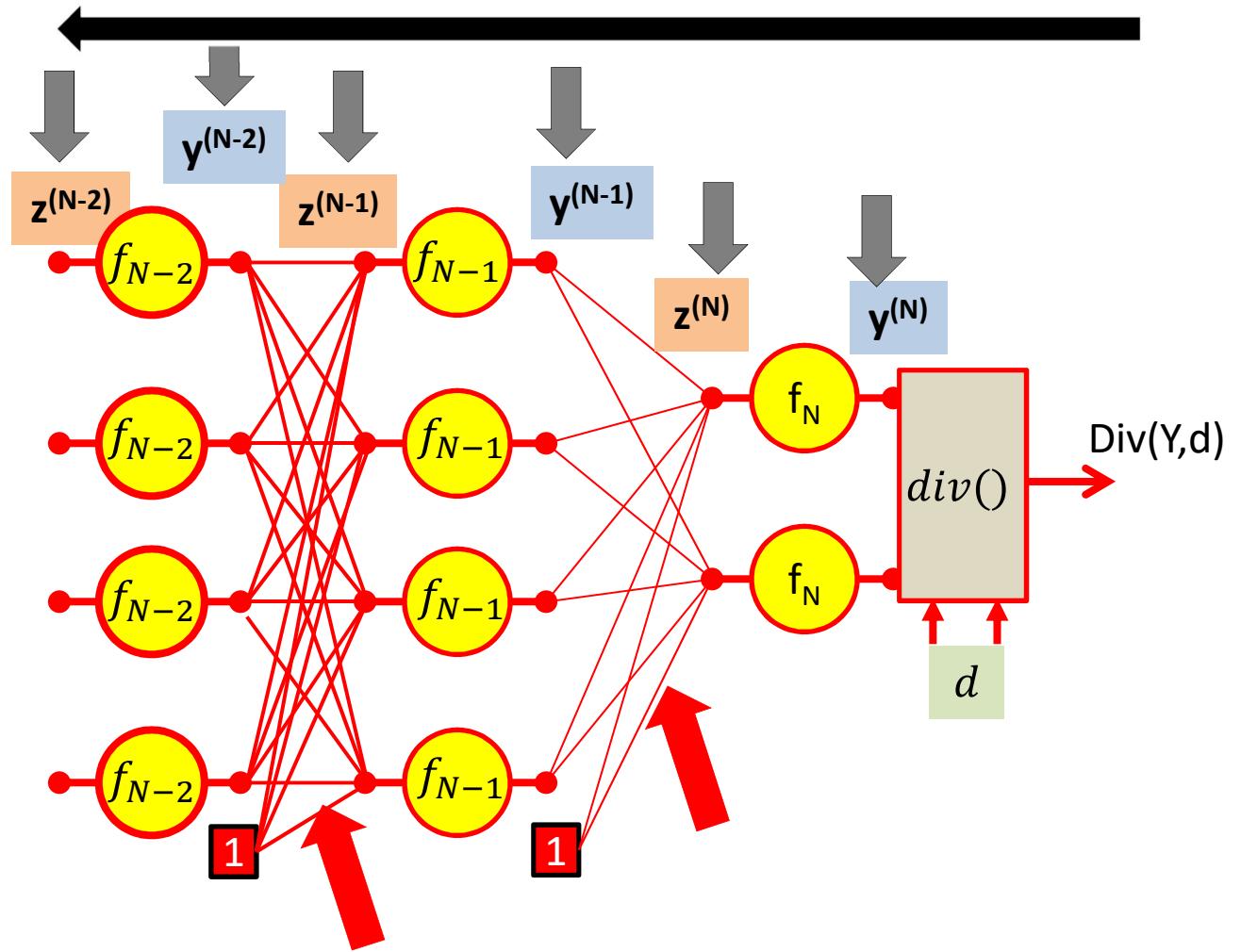


We continue our way backwards in the order shown

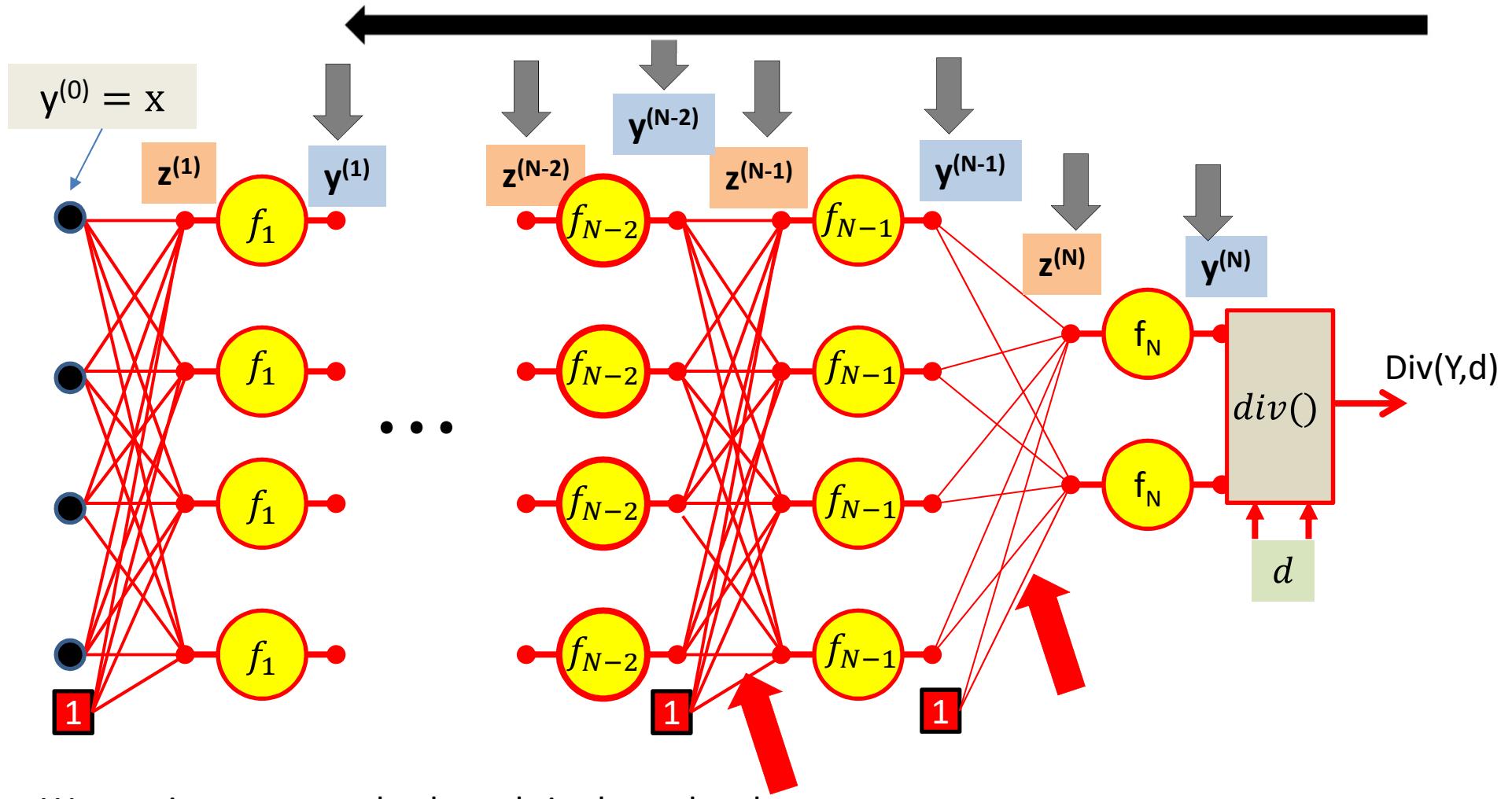
$$\nabla_{Y^{(N-2)}} div(.)$$



...

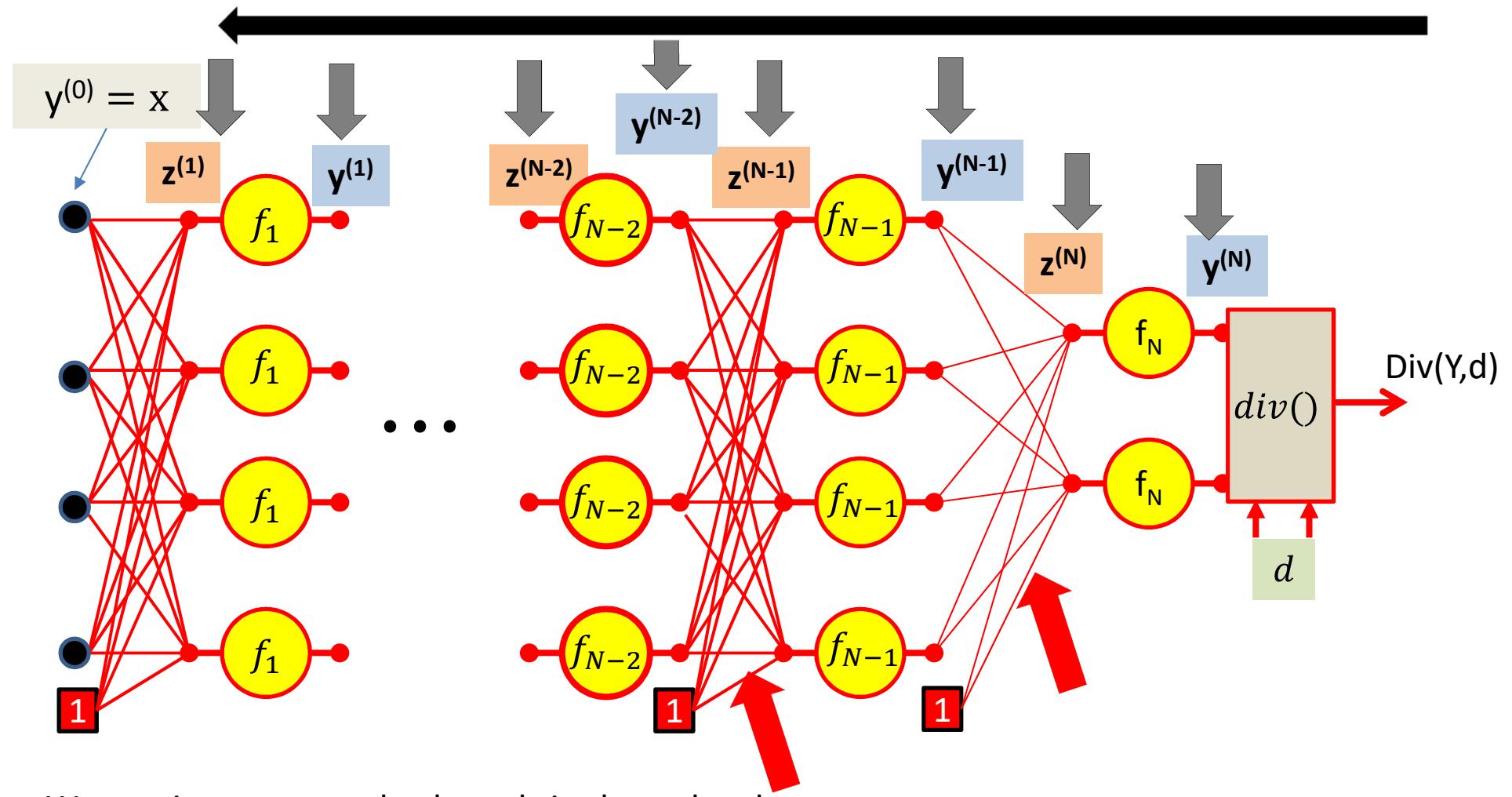


$$\nabla_{z^{(N-2)}} div(.)$$



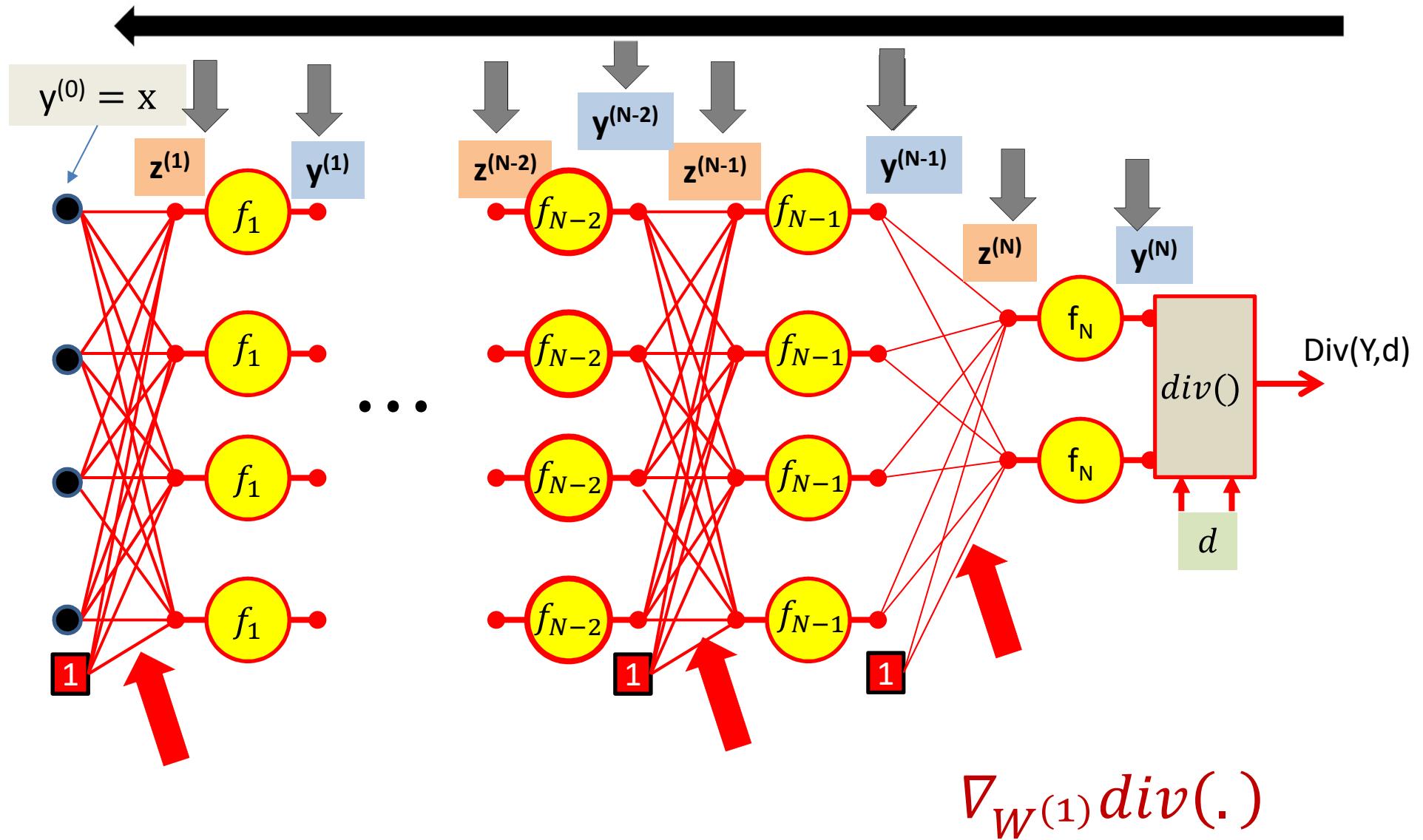
We continue our way backwards in the order shown

$$\nabla_{Y^{(1)}} \text{div}(\cdot)$$



We continue our way backwards in the order shown

$$\nabla_{z^{(1)}} div(.)$$

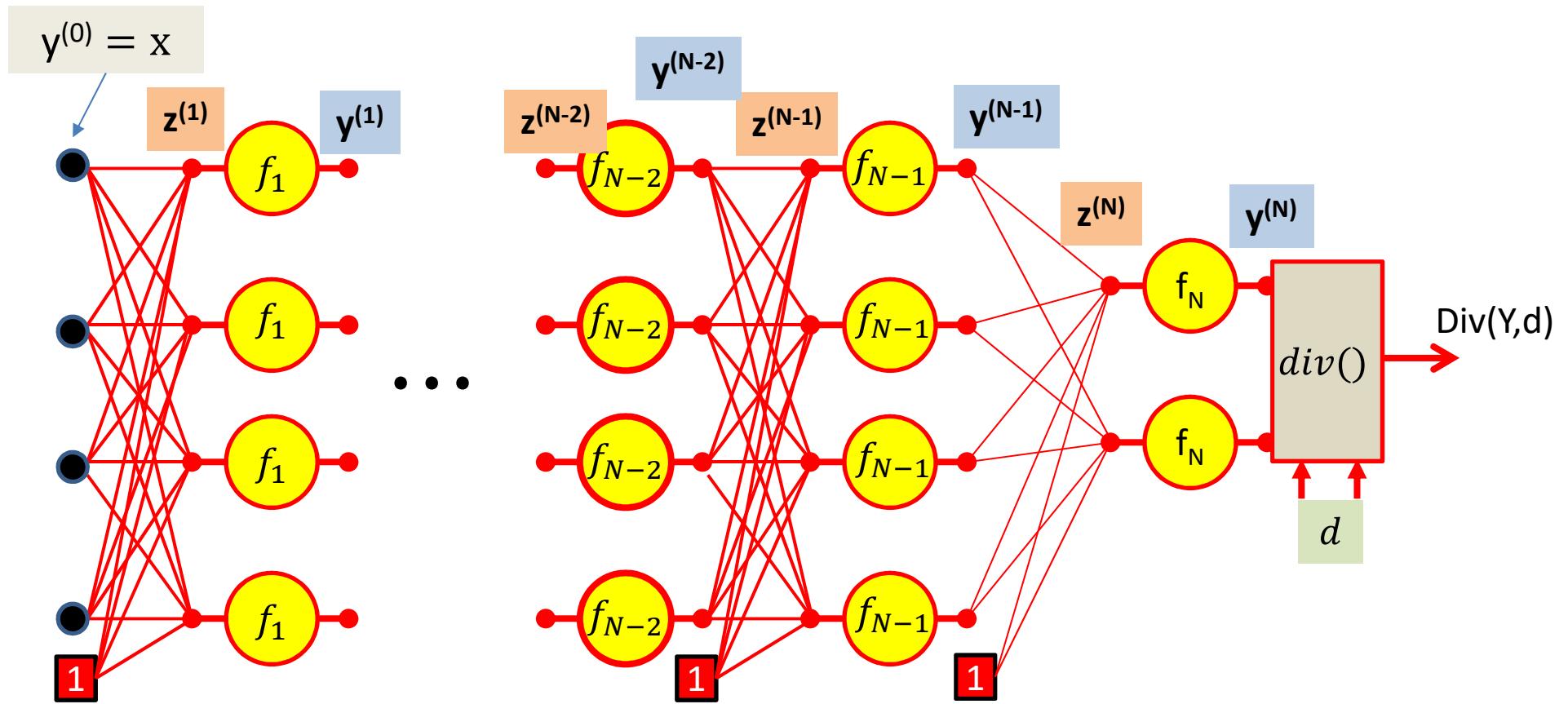


We continue our way backwards in the order shown

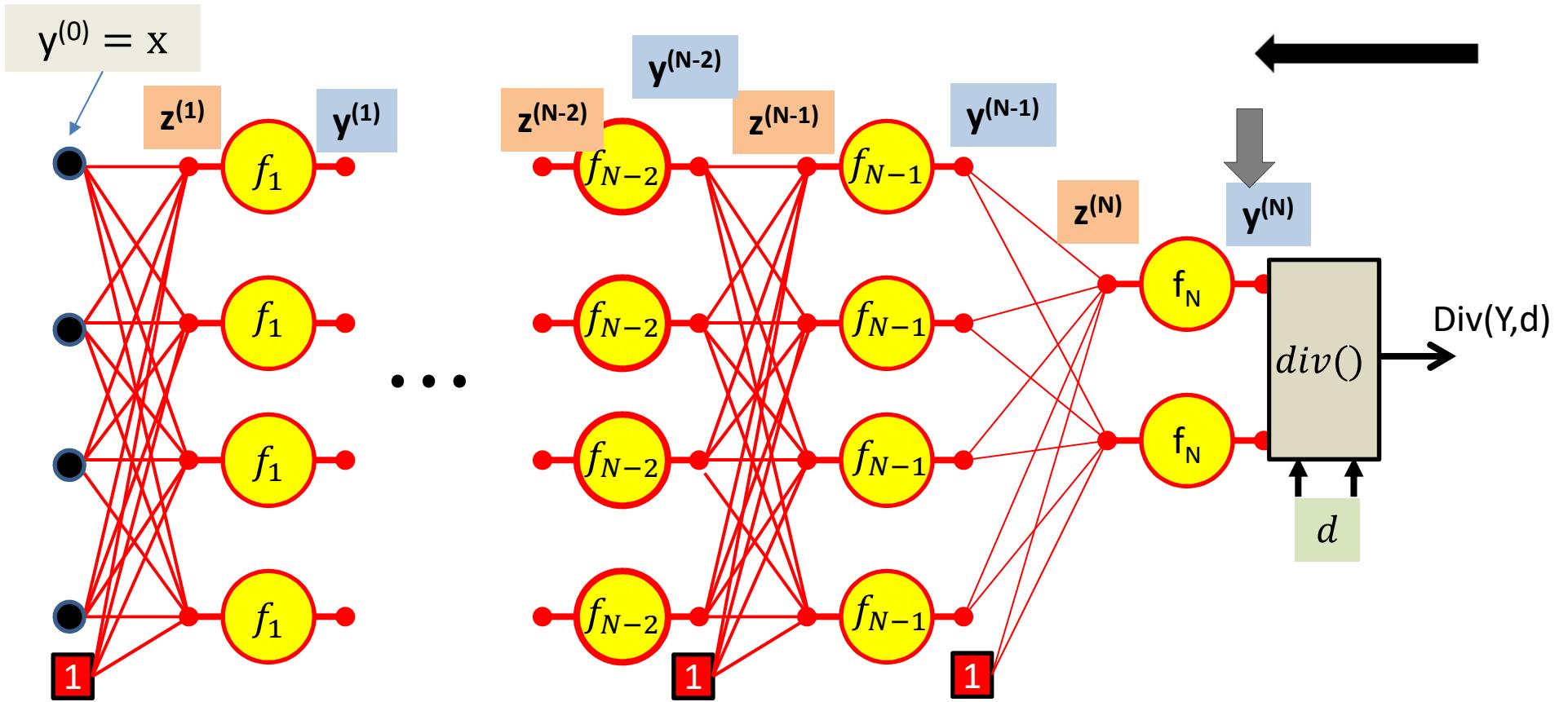
# Backward Gradient Computation

- Lets actually see the math..

# Computing derivatives



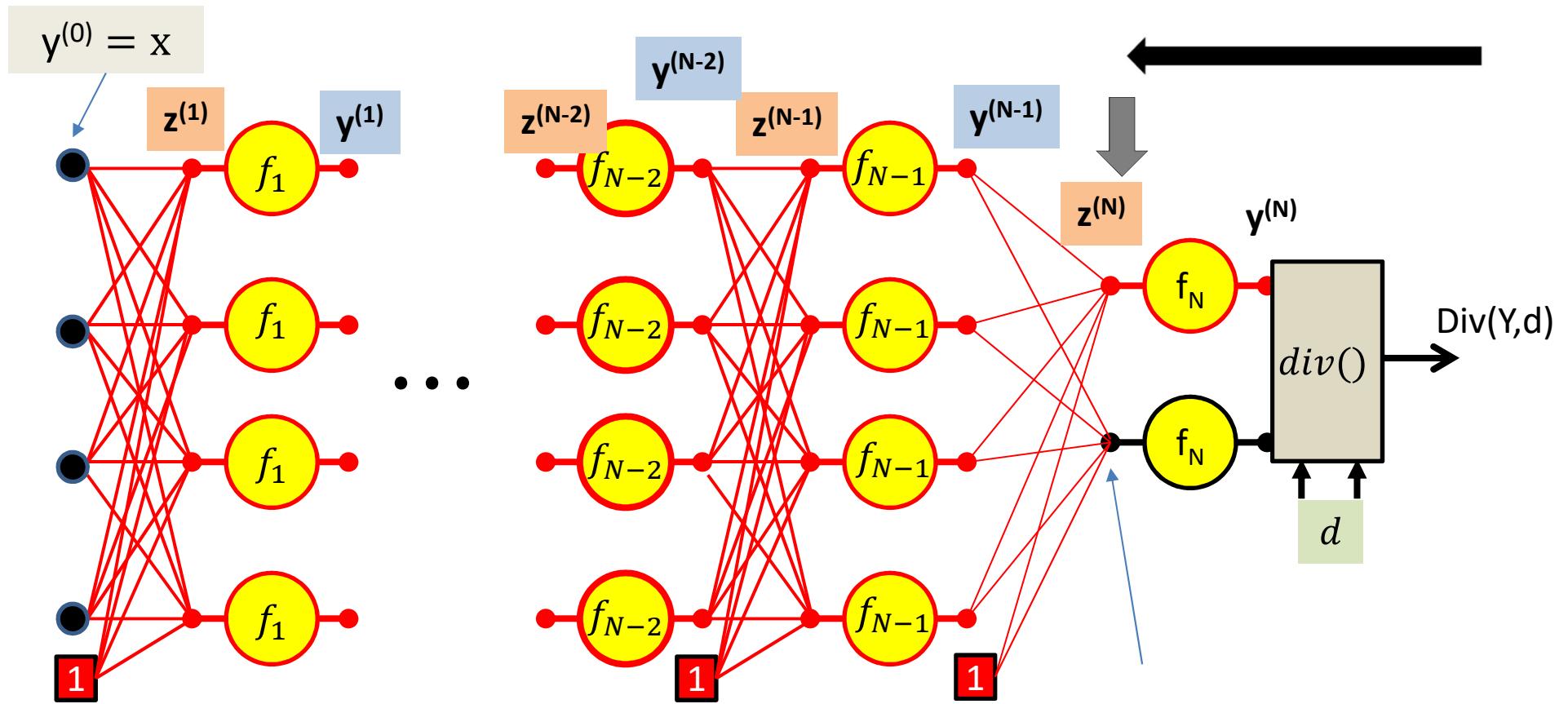
# Computing derivatives



The derivative w.r.t the actual output of the network is simply the derivative w.r.t to the output of the final layer of the network

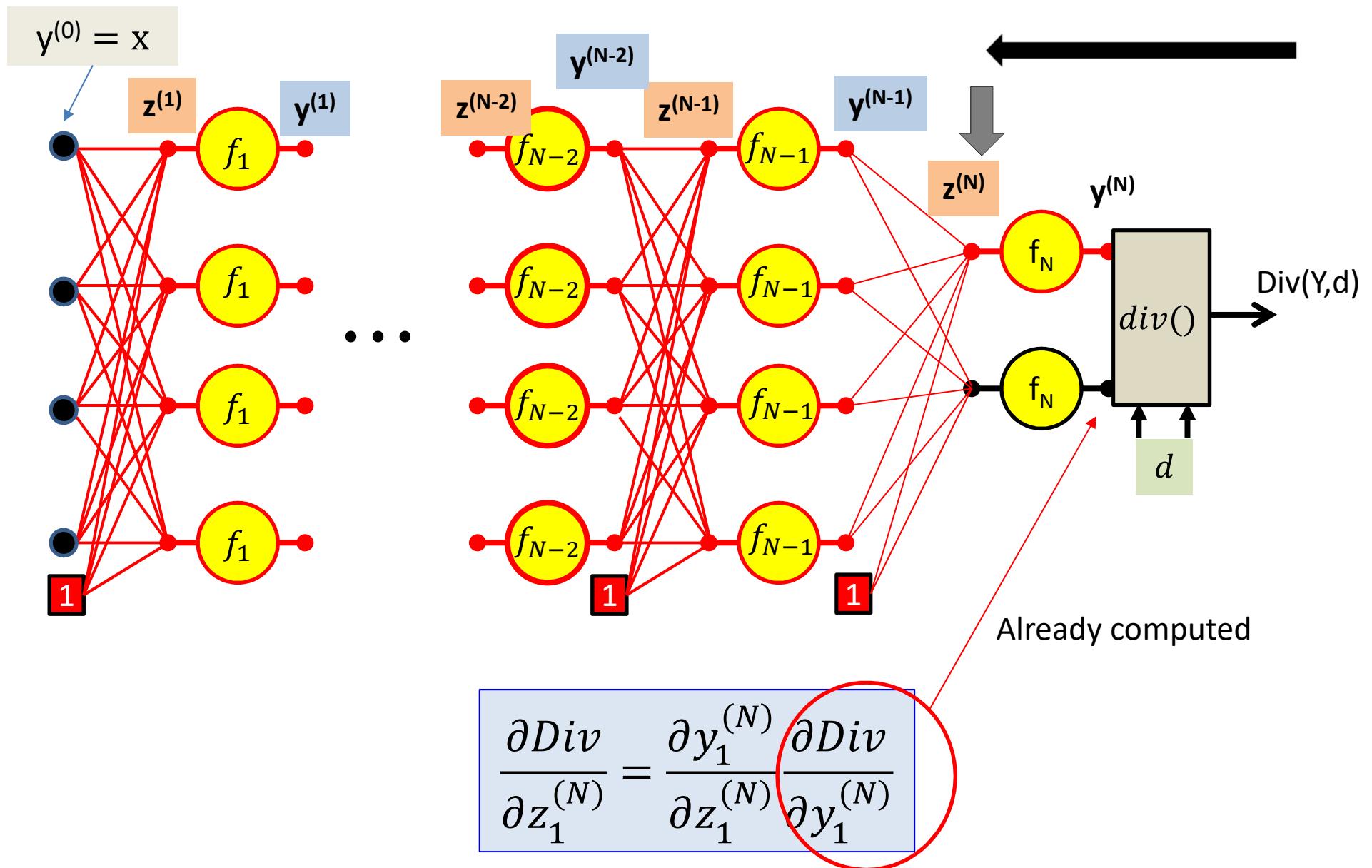
$$\frac{\partial \text{Div}(Y, d)}{\partial y_i} = \frac{\partial \text{Div}(Y, d)}{\partial y_i^{(N)}}$$

# Computing derivatives

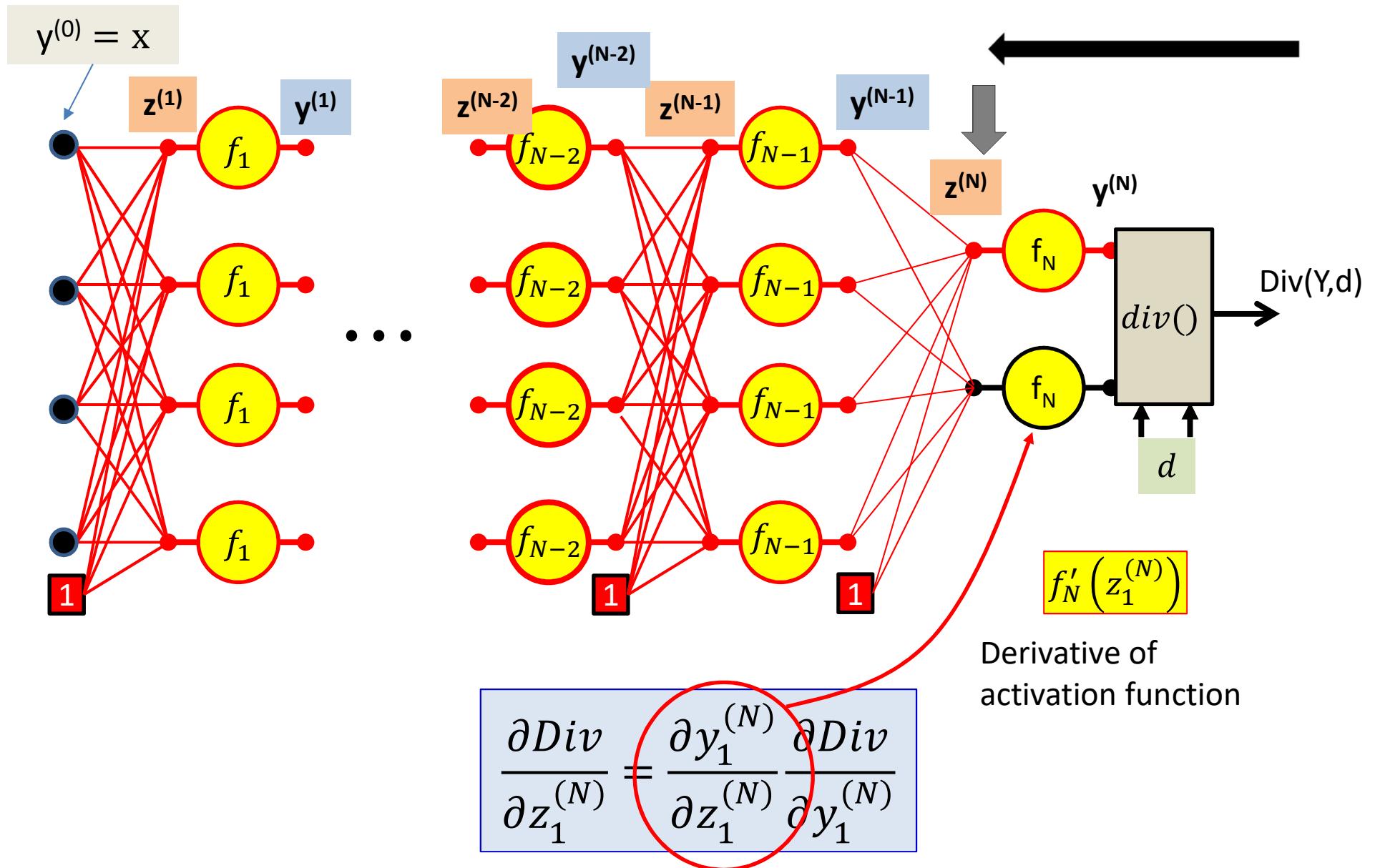


$$\frac{\partial Div}{\partial z_1^{(N)}} = \frac{\partial y_1^{(N)}}{\partial z_1^{(N)}} \frac{\partial Div}{\partial y_1^{(N)}}$$

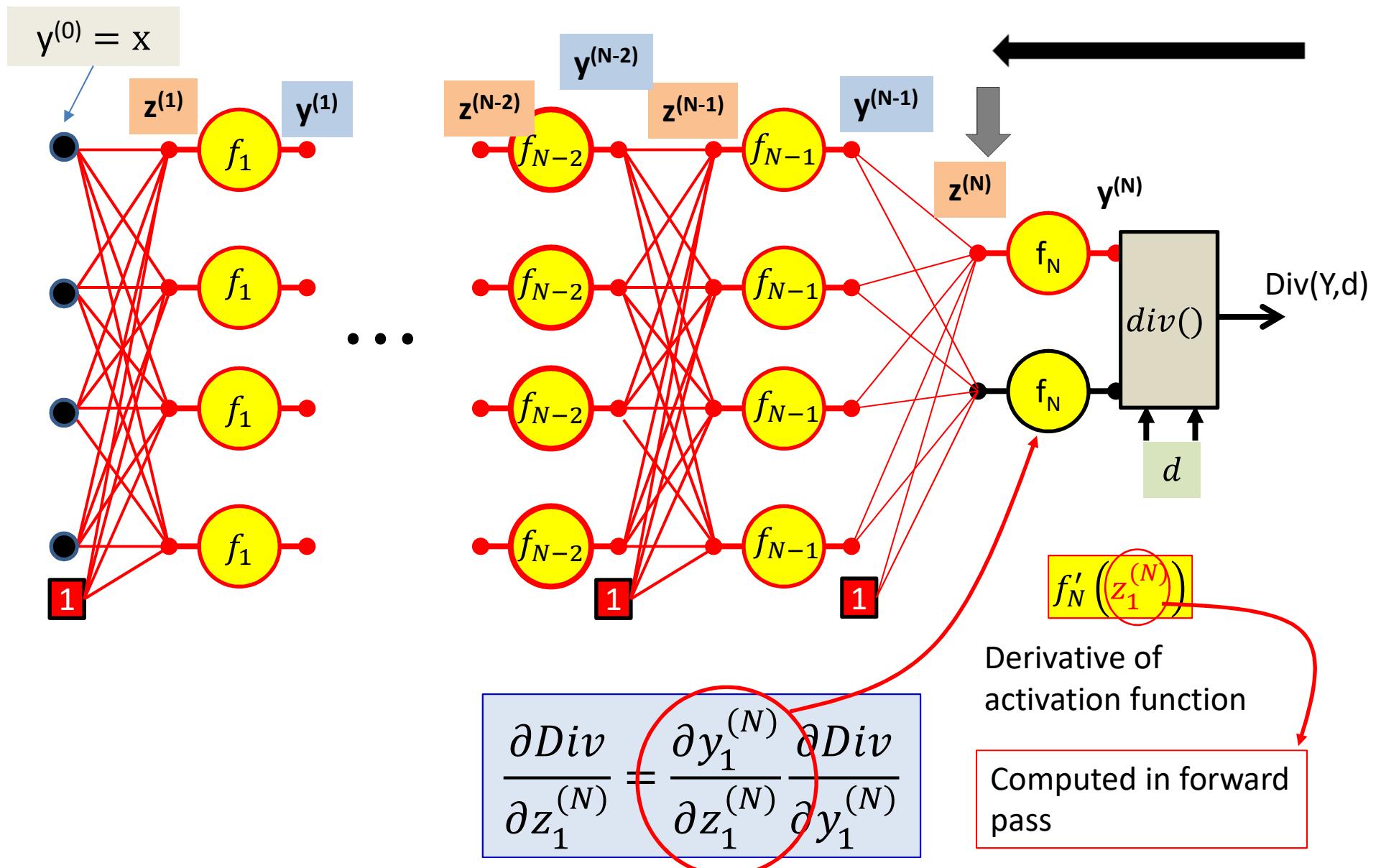
# Computing derivatives



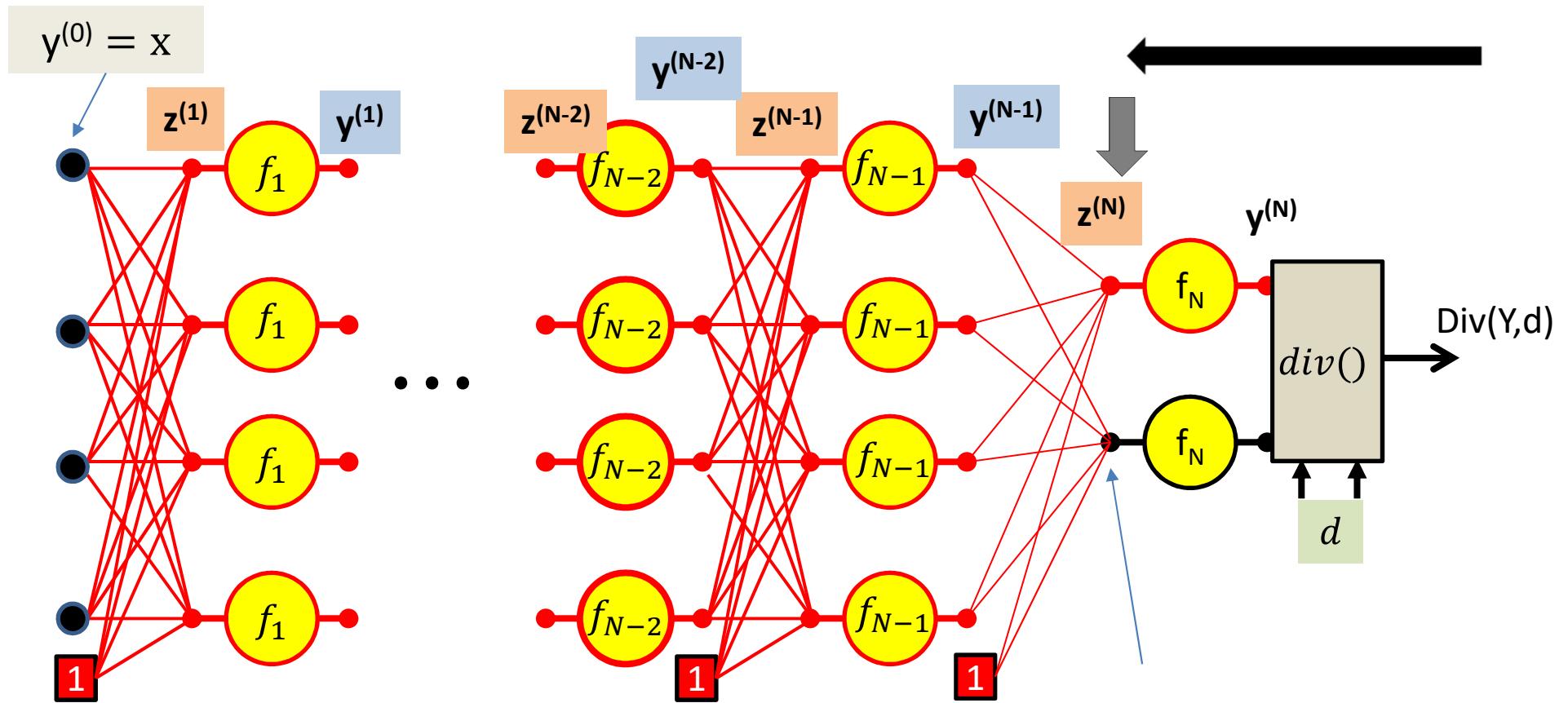
# Computing derivatives



# Computing derivatives

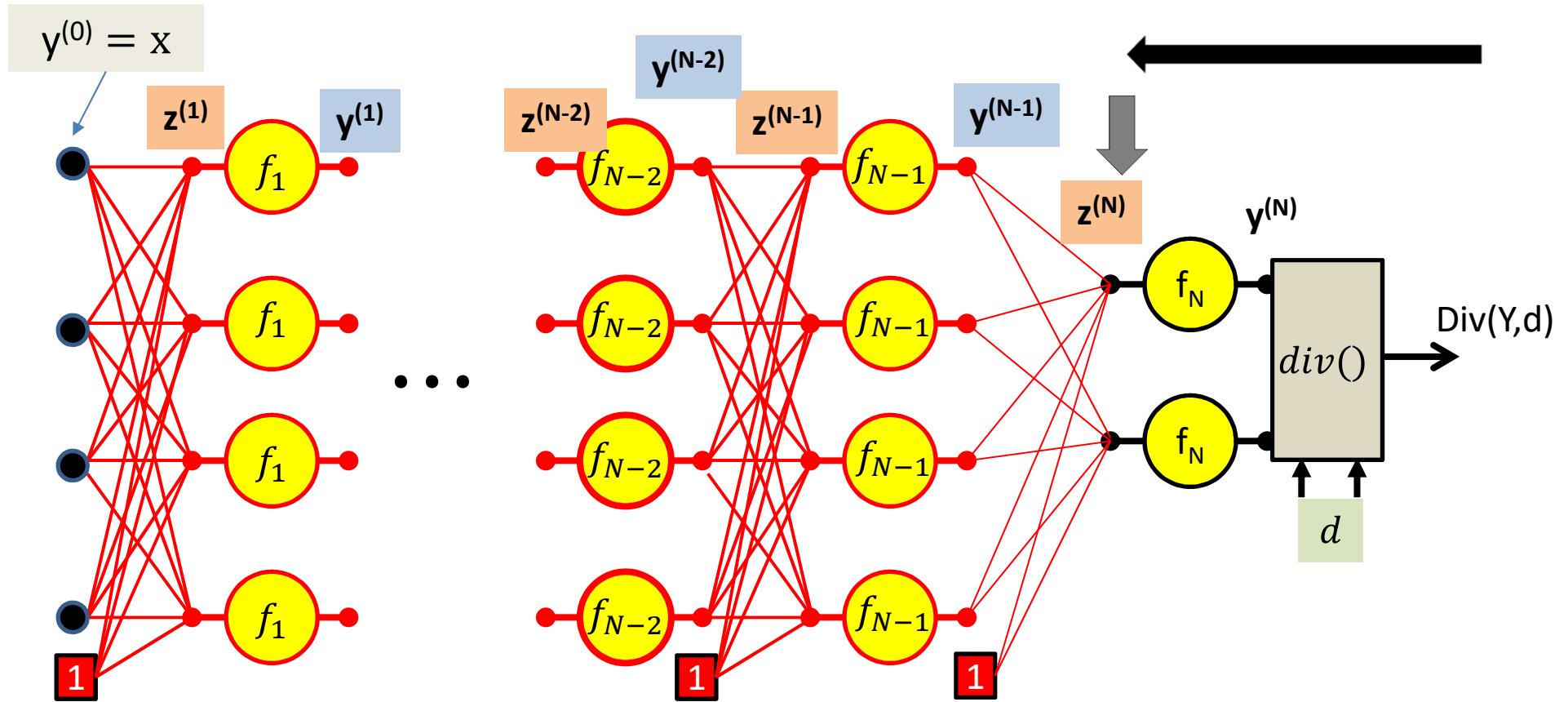


# Computing derivatives



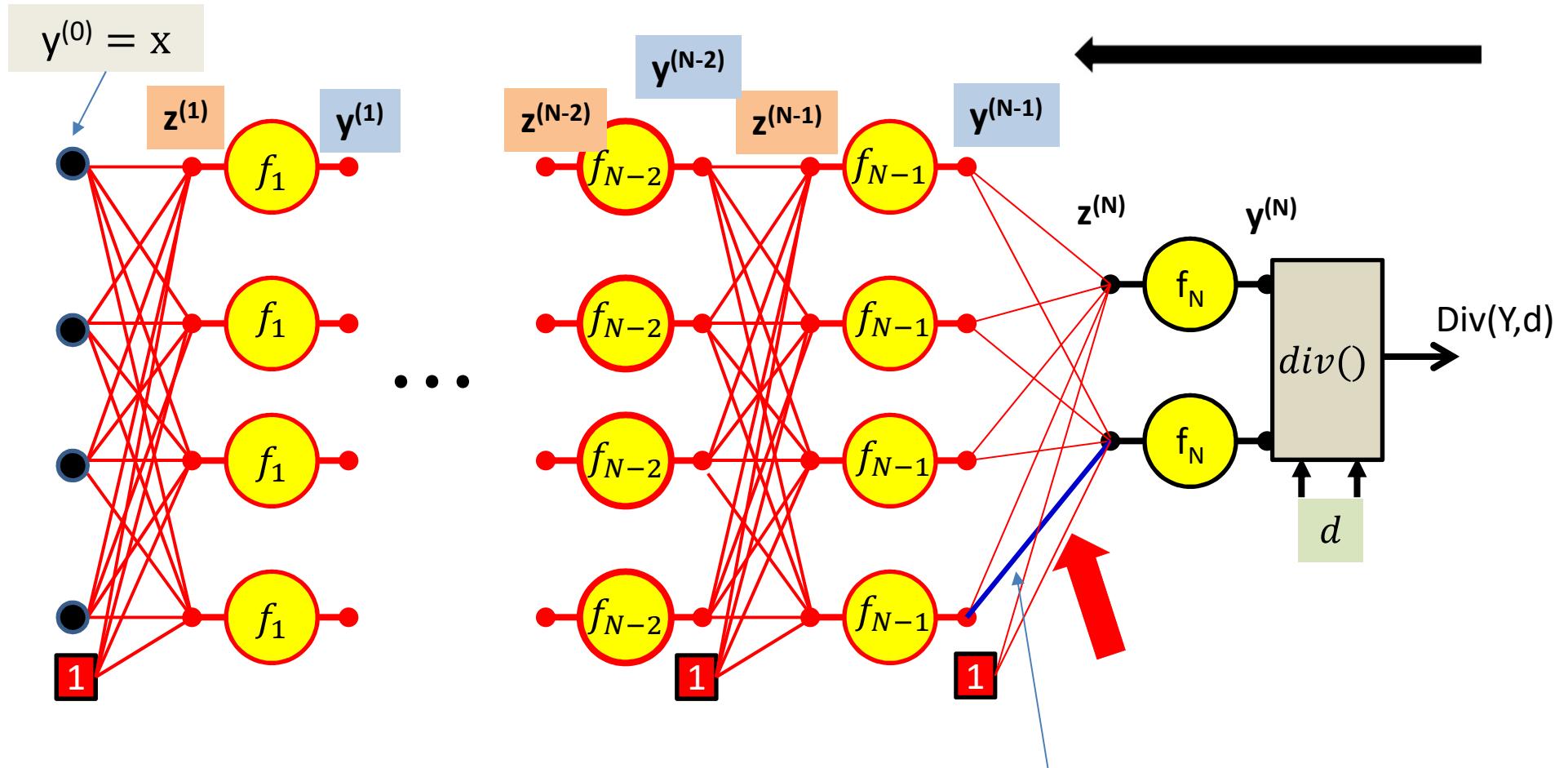
$$\frac{\partial Div}{\partial z_1^{(N)}} = f'_N(z_1^{(N)}) \frac{\partial Div}{\partial y_1^{(N)}}$$

# Computing derivatives



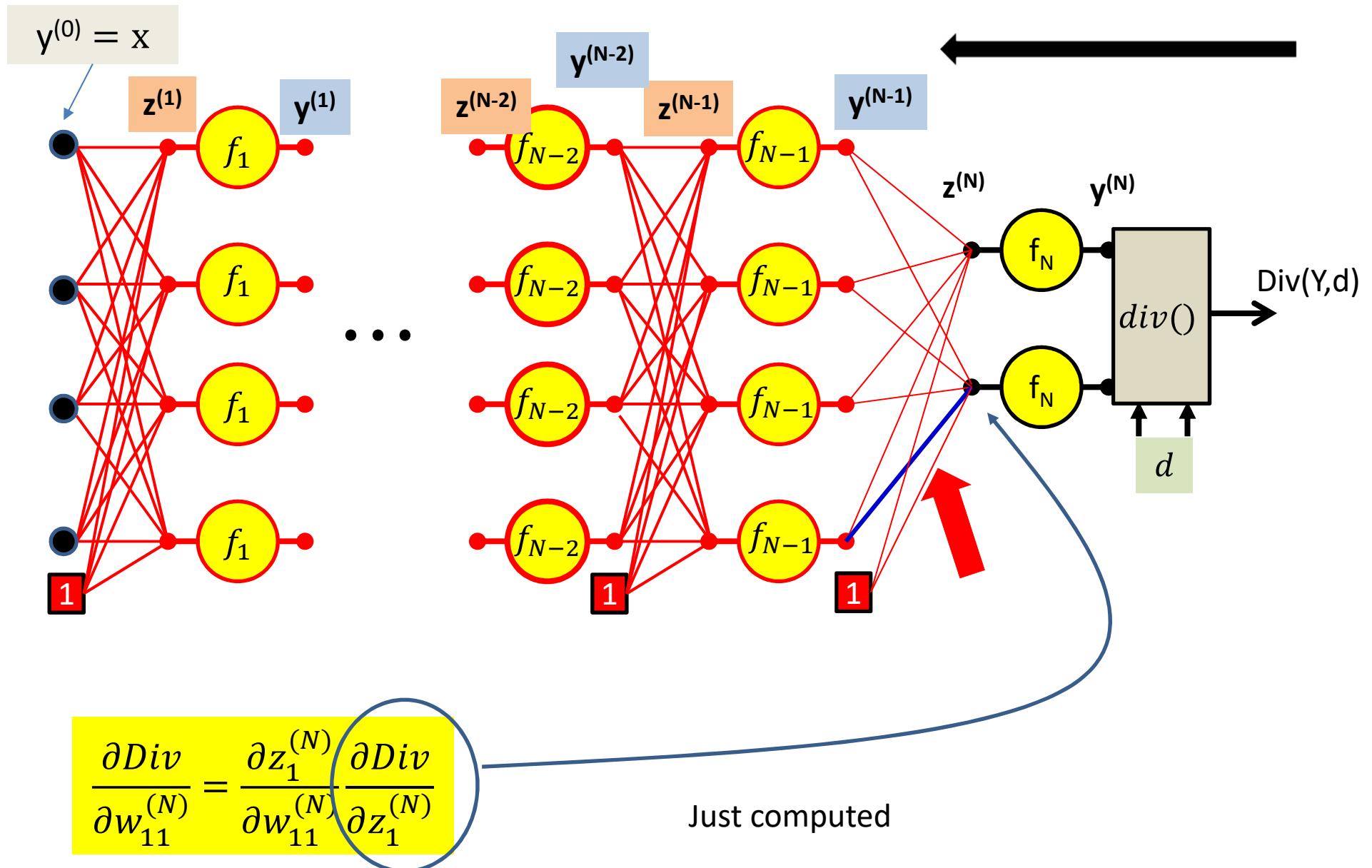
$$\frac{\partial Div}{\partial z_i^{(N)}} = f'_N(z_i^{(N)}) \frac{\partial Div}{\partial y_i^{(N)}}$$

# Computing derivatives

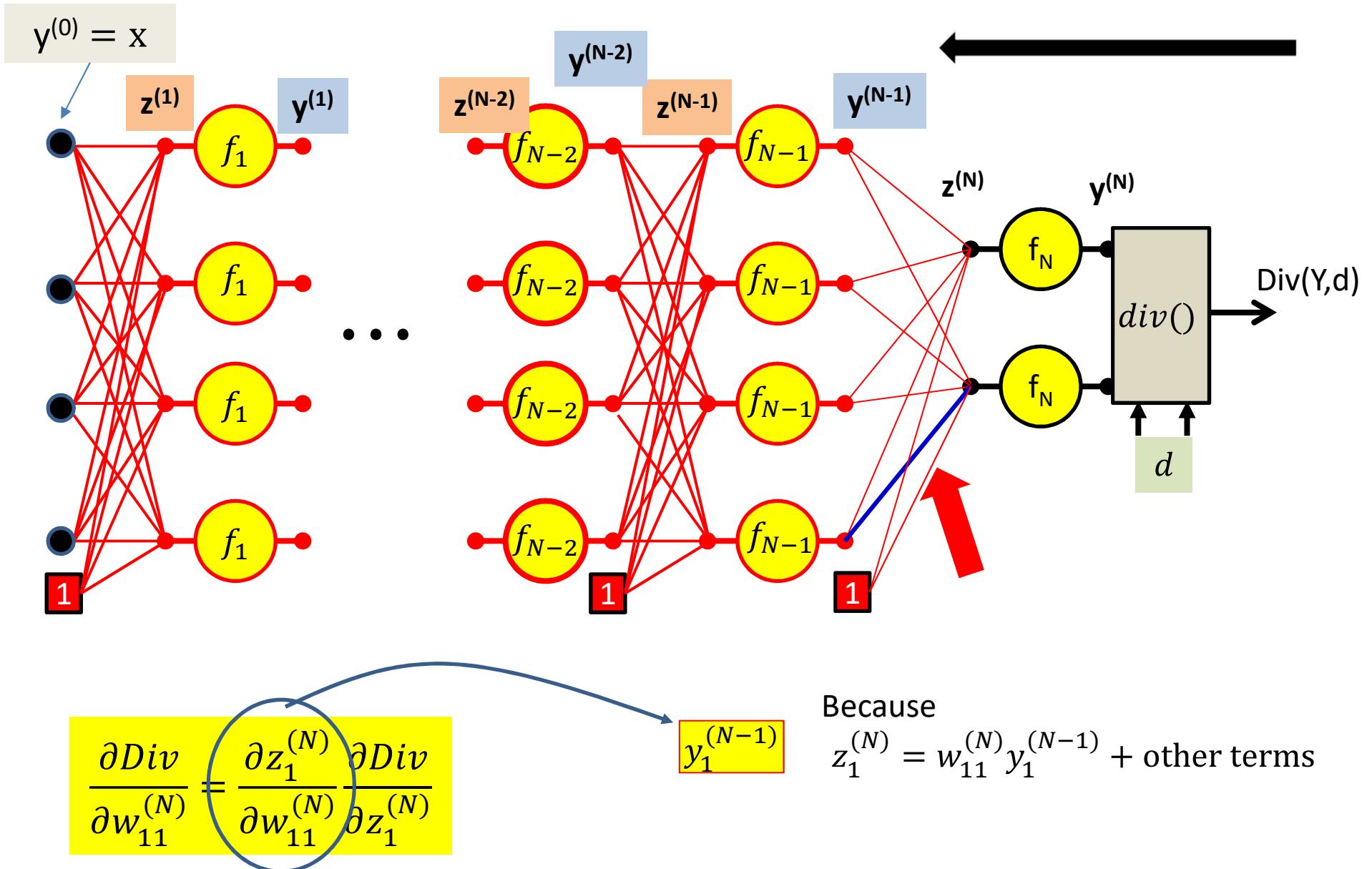


$$\frac{\partial \text{Div}}{\partial w_{11}^{(N)}} = \frac{\partial z_1^{(N)}}{\partial w_{11}^{(N)}} \frac{\partial \text{Div}}{\partial z_1^{(N)}}$$

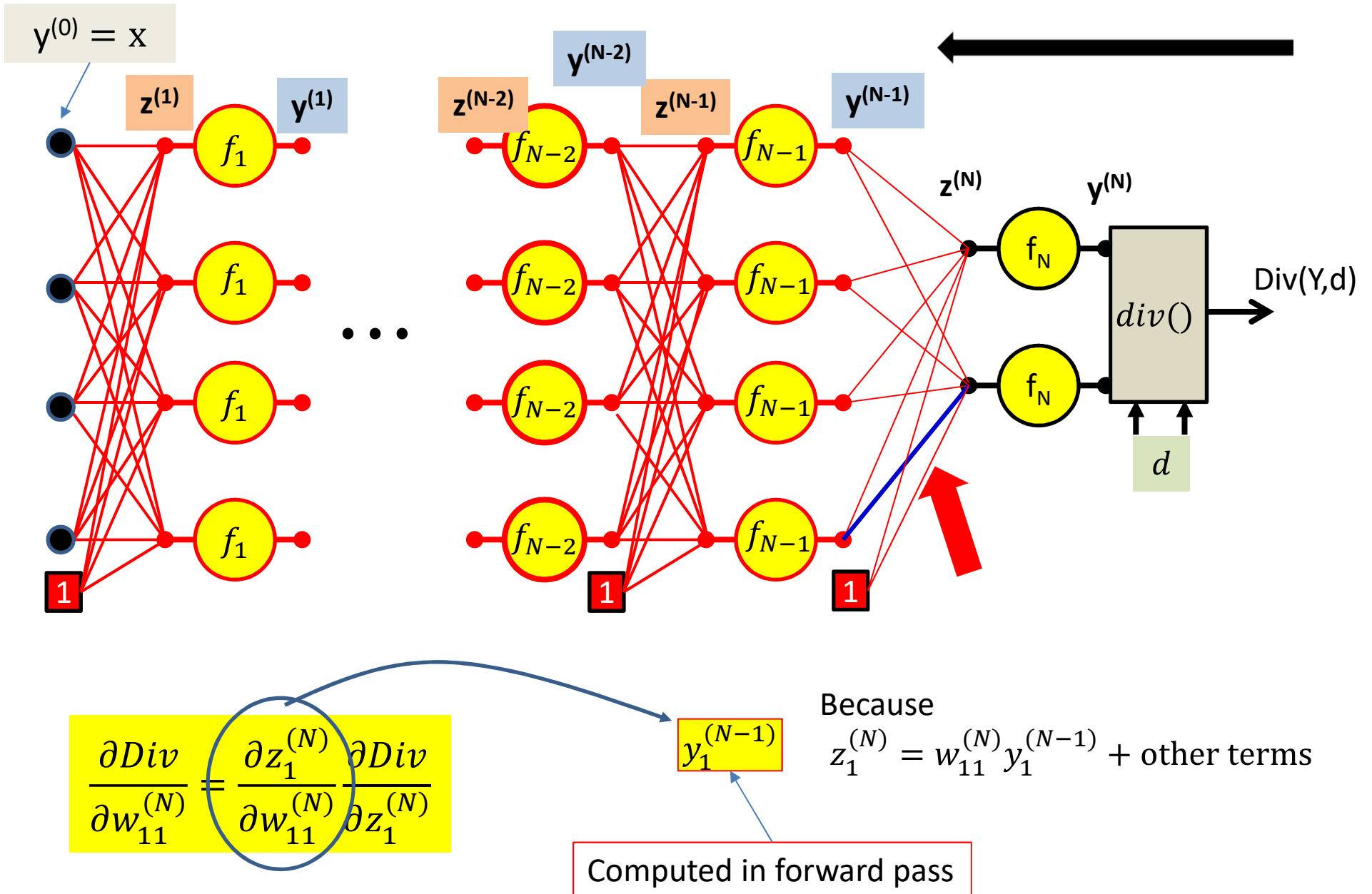
# Computing derivatives



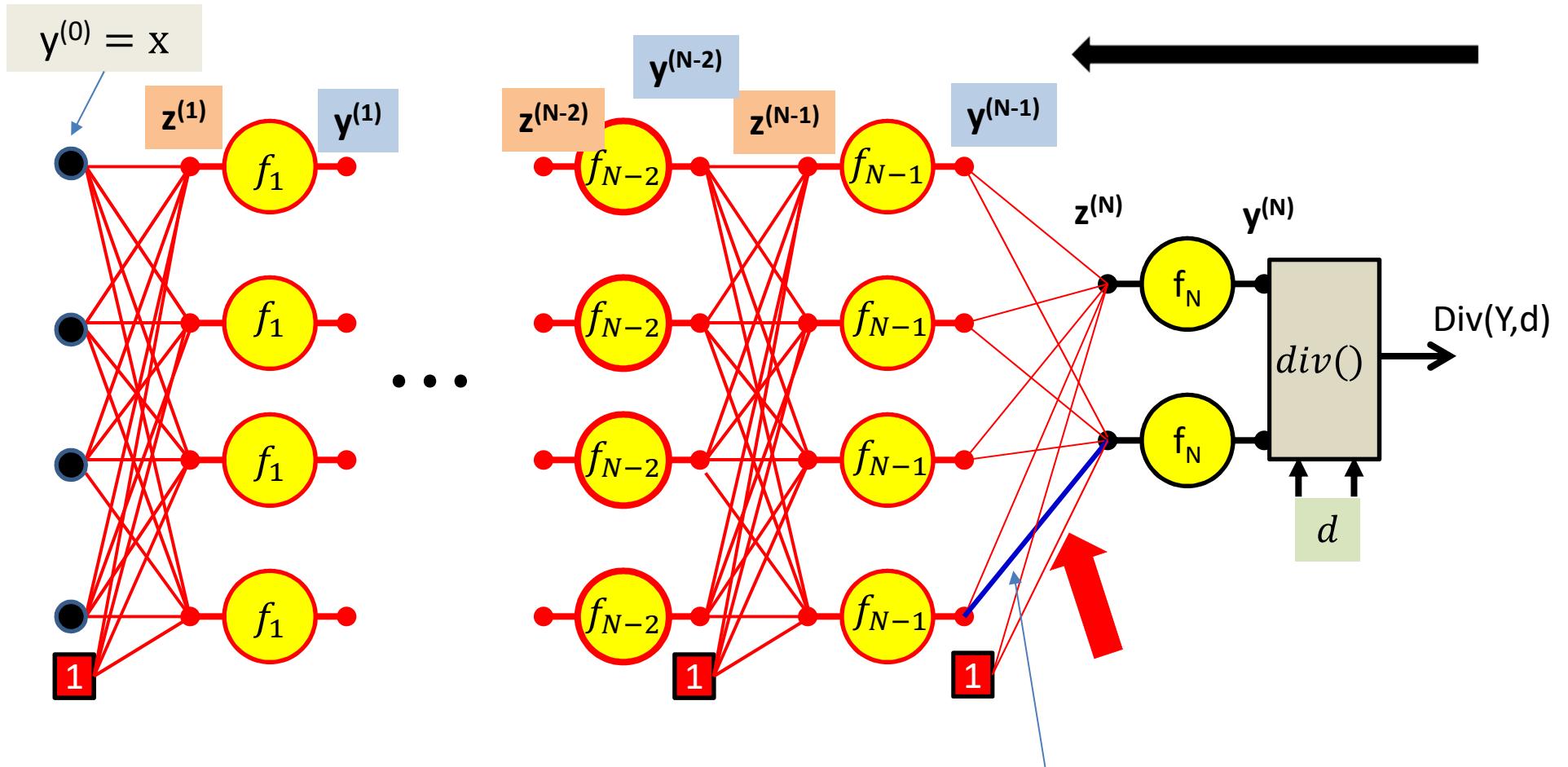
# Computing derivatives



# Computing derivatives

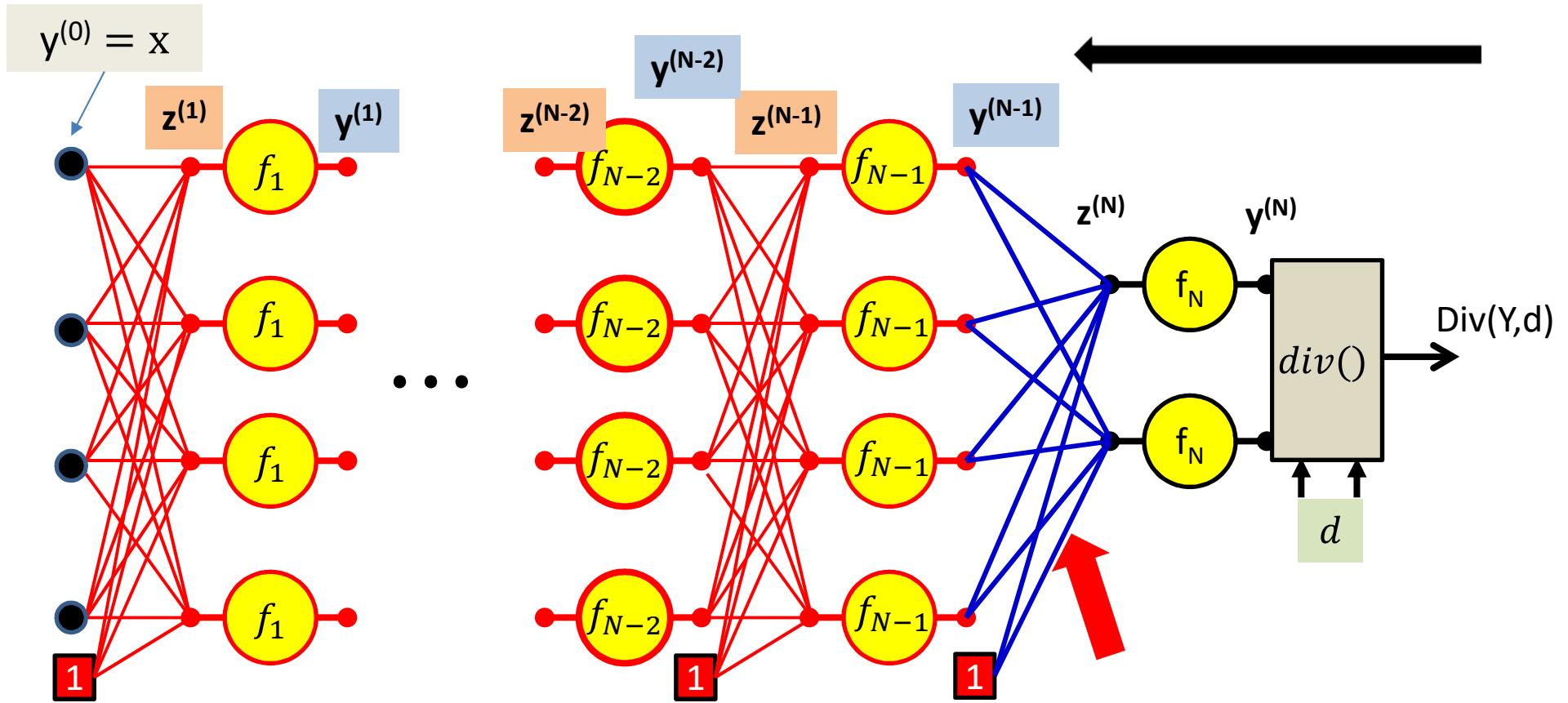


# Computing derivatives



$$\frac{\partial \text{Div}}{\partial w_{11}^{(N)}} = y_1^{(N-1)} \frac{\partial \text{Div}}{\partial z_1^{(N)}}$$

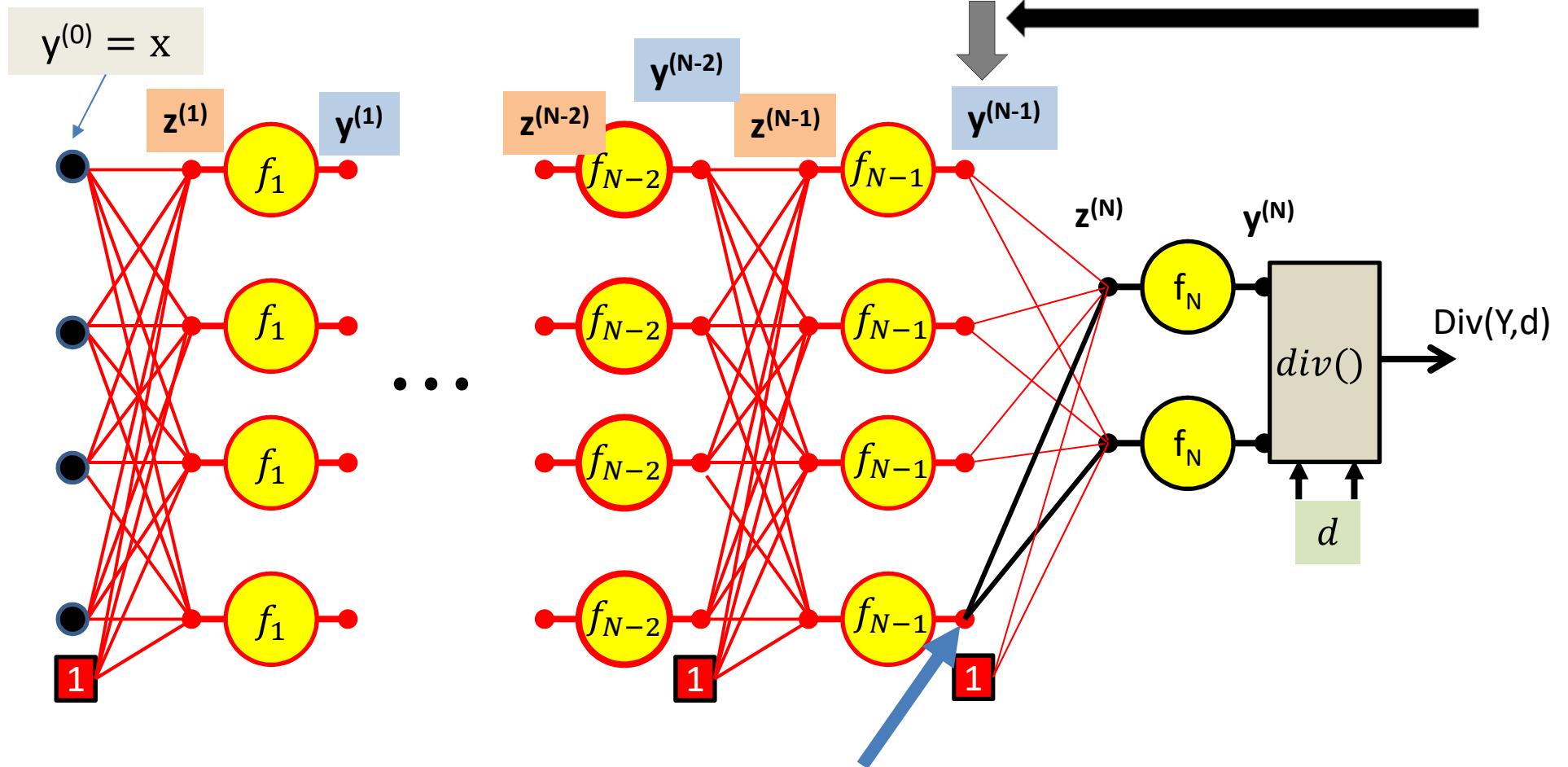
# Computing derivatives



$$\frac{\partial Div}{\partial w_{ij}^{(N)}} = y_i^{(N-1)} \frac{\partial Div}{\partial z_j^{(N)}}$$

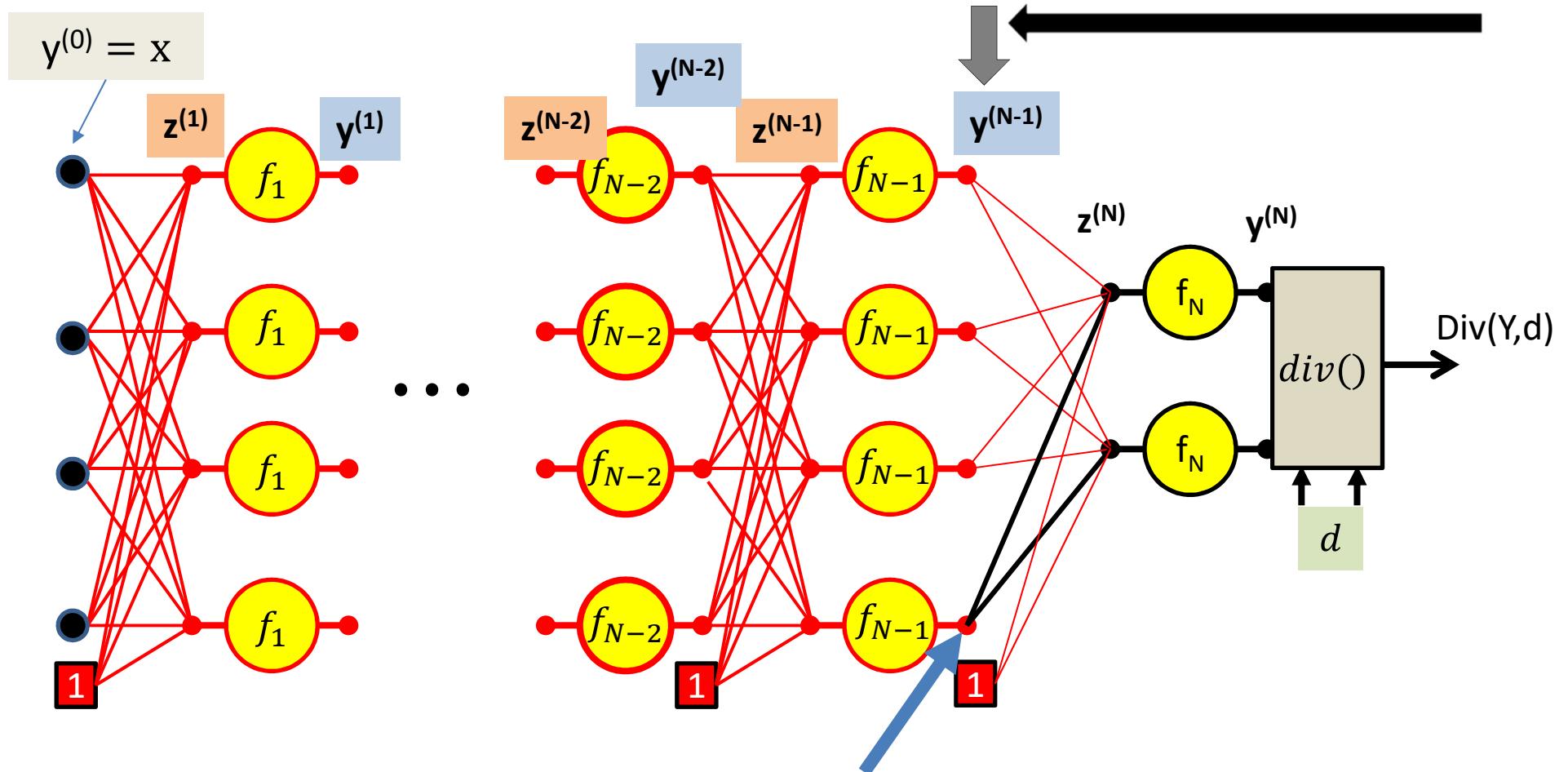
For the bias term  $y_0^{(N-1)} = 1$

# Computing derivatives



$$\frac{\partial Div}{\partial y_1^{(N-1)}} = \sum_j \frac{\partial z_j^{(N)}}{\partial y_1^{(N-1)}} \frac{\partial Div}{\partial z_j^{(N)}}$$

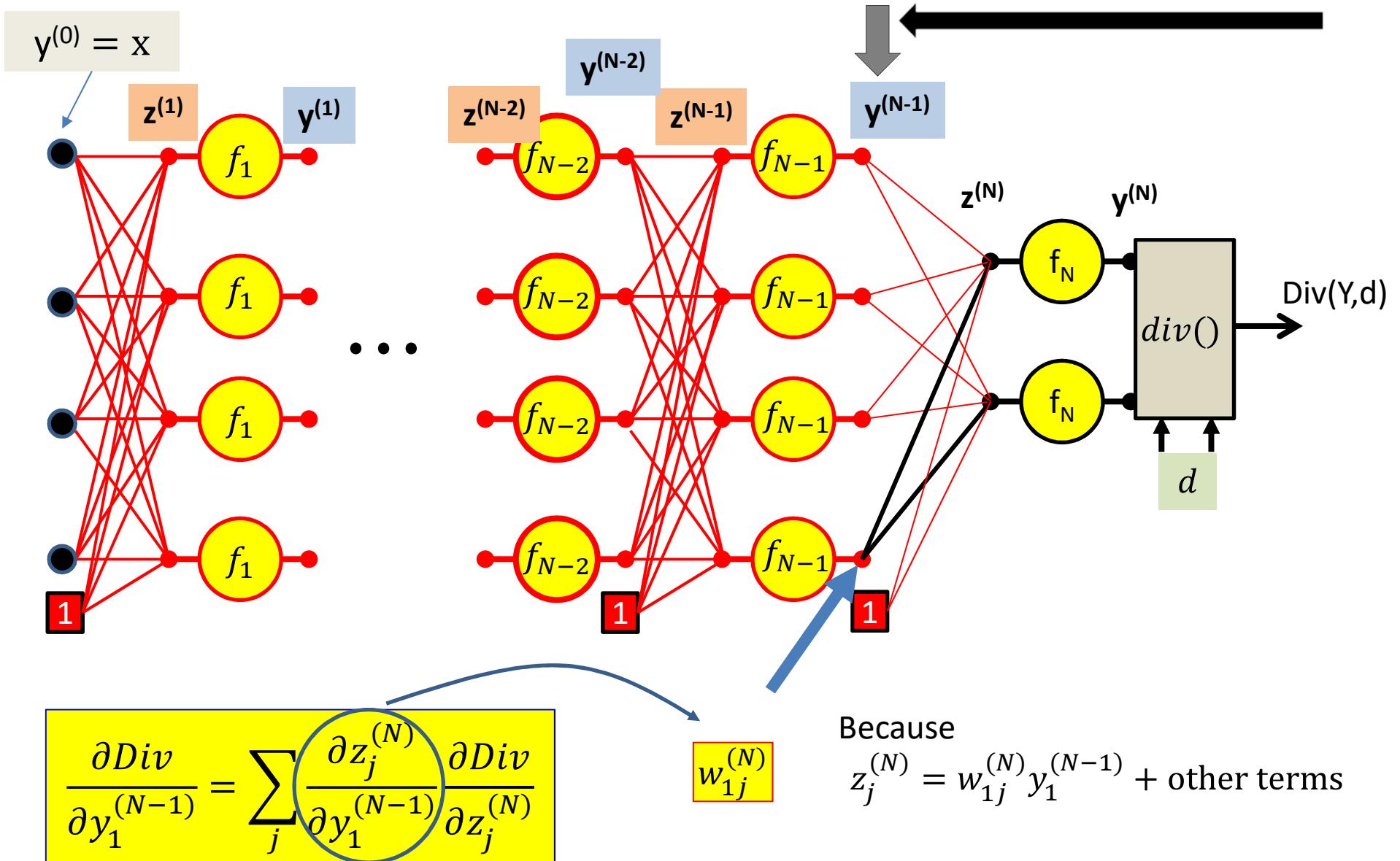
# Computing derivatives



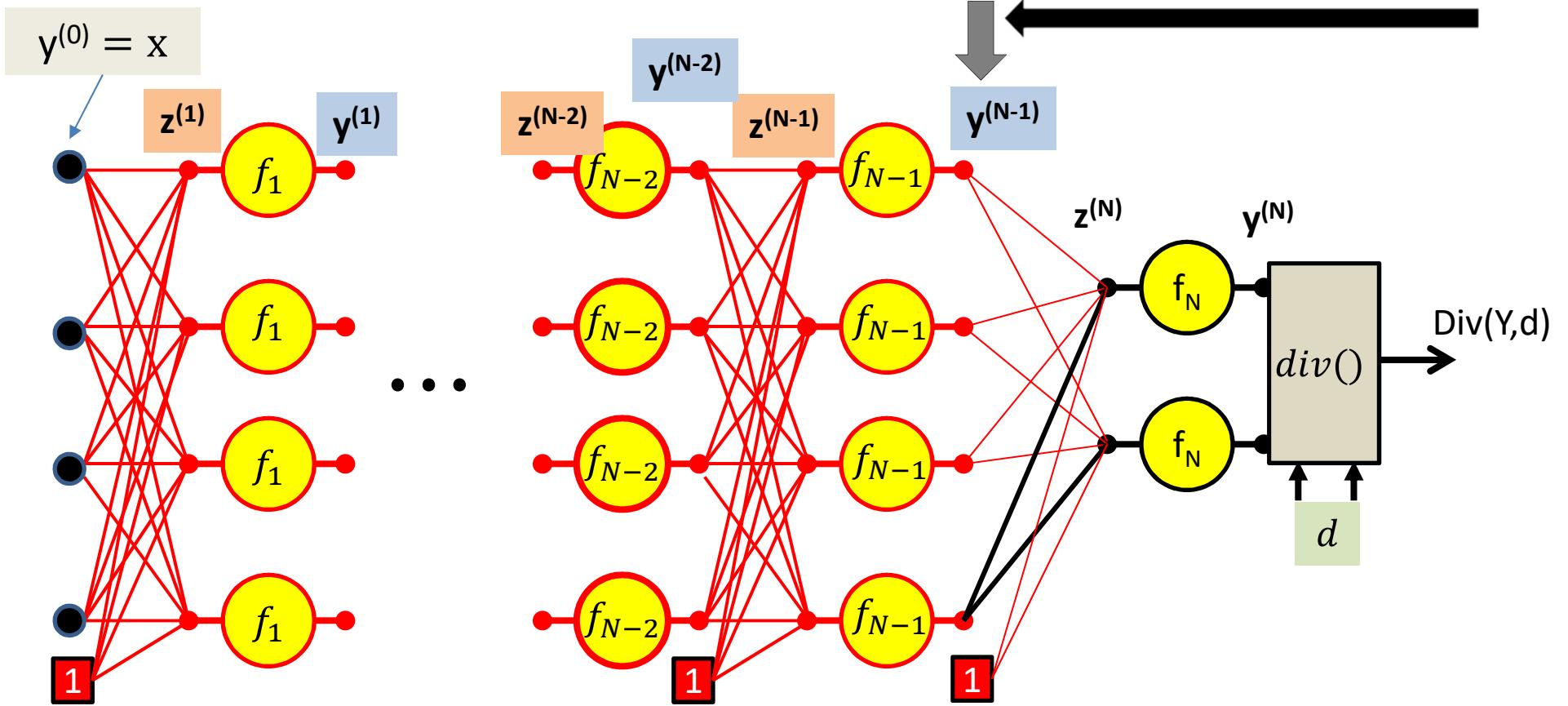
$$\frac{\partial Div}{\partial y_1^{(N-1)}} = \sum_j \frac{\partial z_j^{(N)}}{\partial y_1^{(N-1)}} \frac{\partial Div}{\partial z_j^{(N)}}$$

Already computed

# Computing derivatives

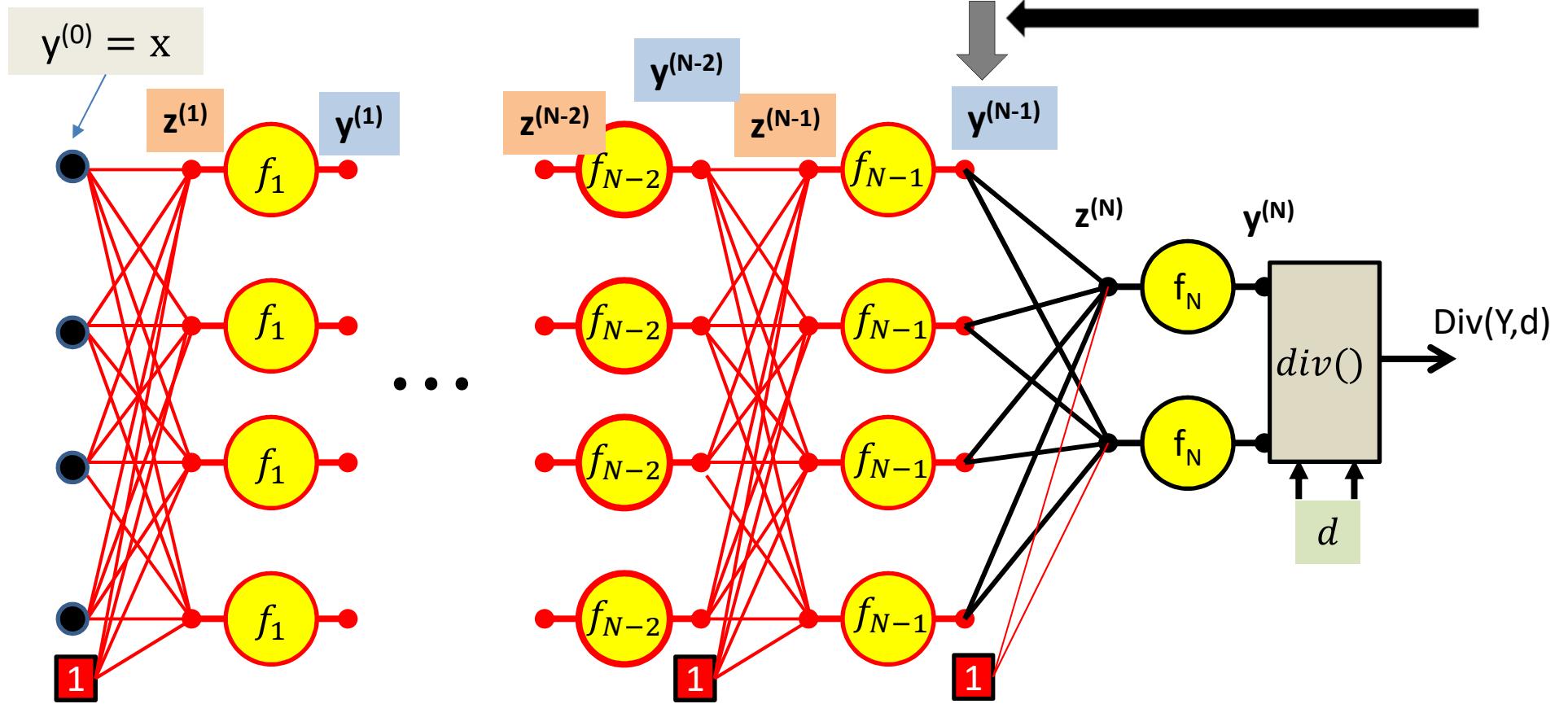


# Computing derivatives



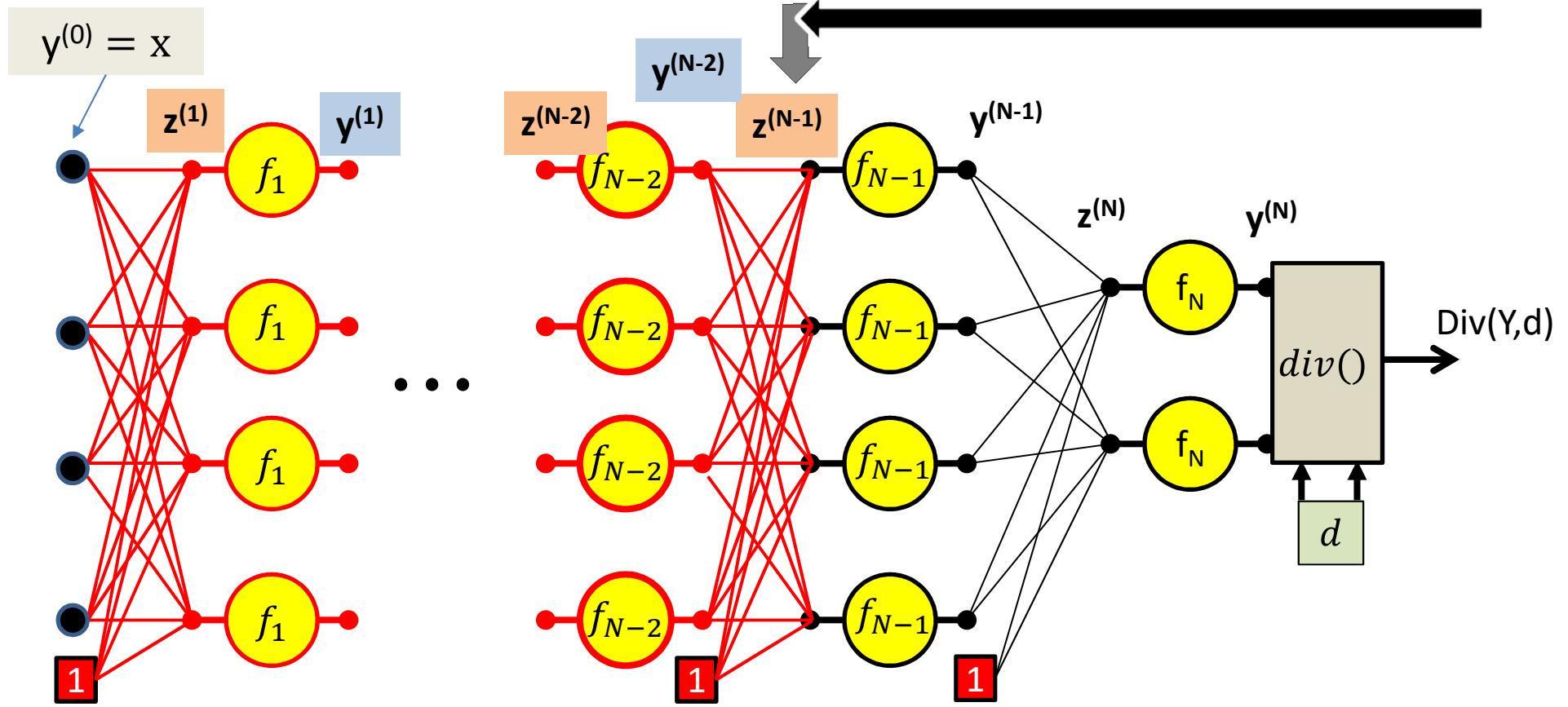
$$\frac{\partial Div}{\partial y_1^{(N-1)}} = \sum_j w_{1j}^{(N)} \frac{\partial Div}{\partial z_j^{(N)}}$$

# Computing derivatives



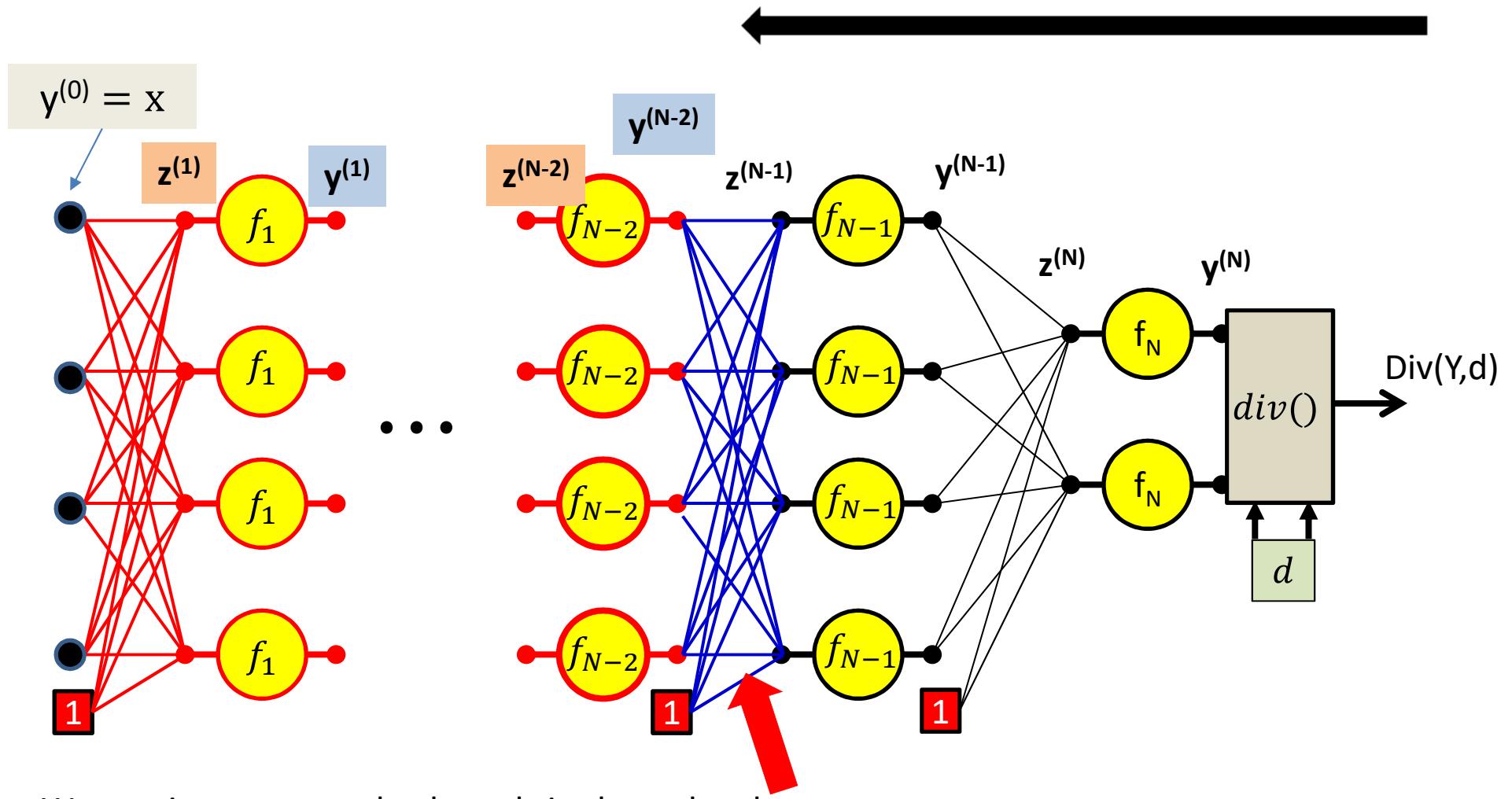
$$\frac{\partial \text{Div}}{\partial y_i^{(N-1)}} = \sum_j w_{ij}^{(N)} \frac{\partial \text{Div}}{\partial z_j^{(N)}}$$

# Computing derivatives



We continue our way backwards in the order shown

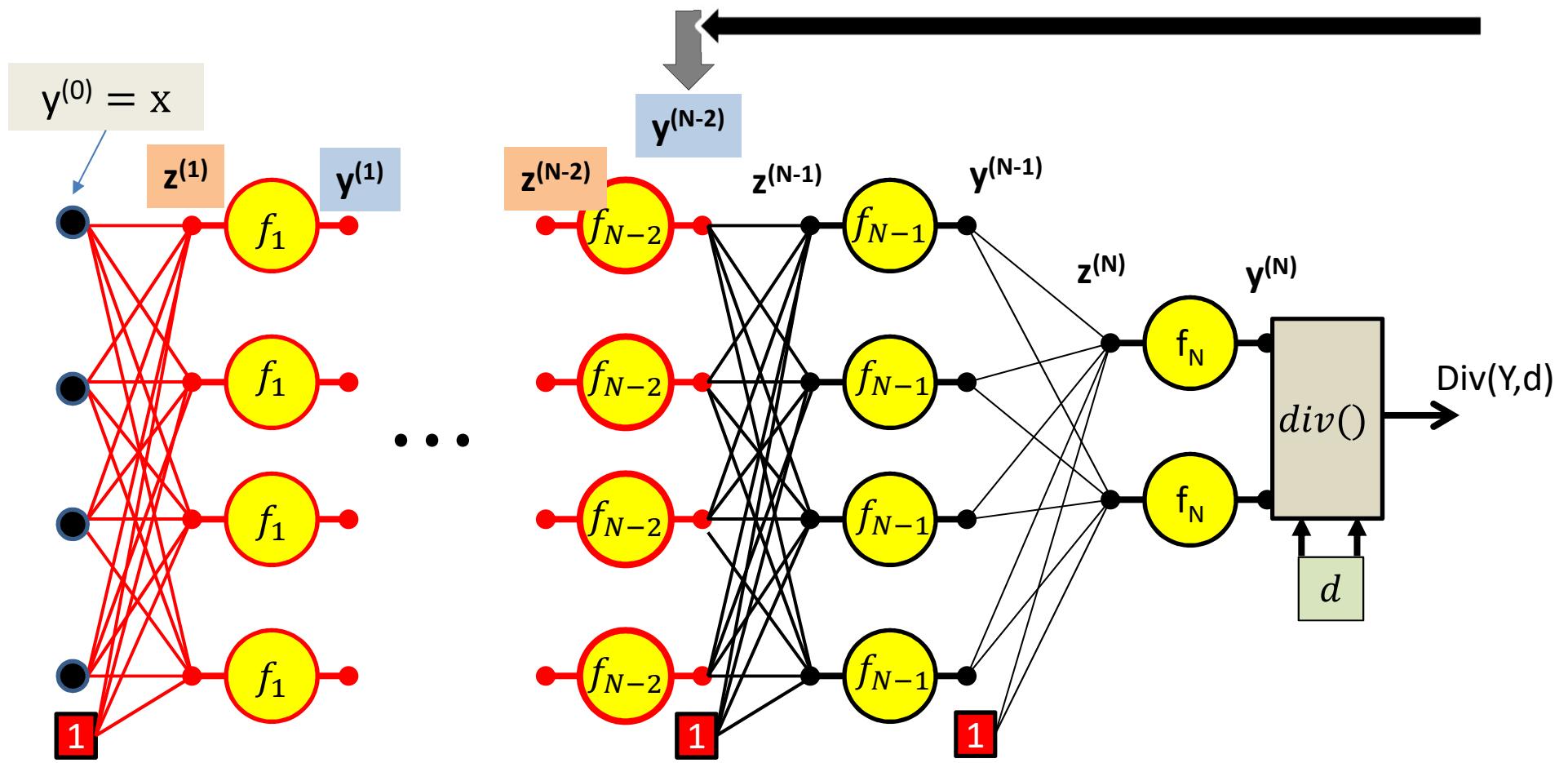
$$\frac{\partial Div}{\partial z_i^{(N-1)}} = f'_{N-1}(z_i^{(N-1)}) \frac{\partial Div}{\partial y_i^{(N-1)}}$$



We continue our way backwards in the order shown

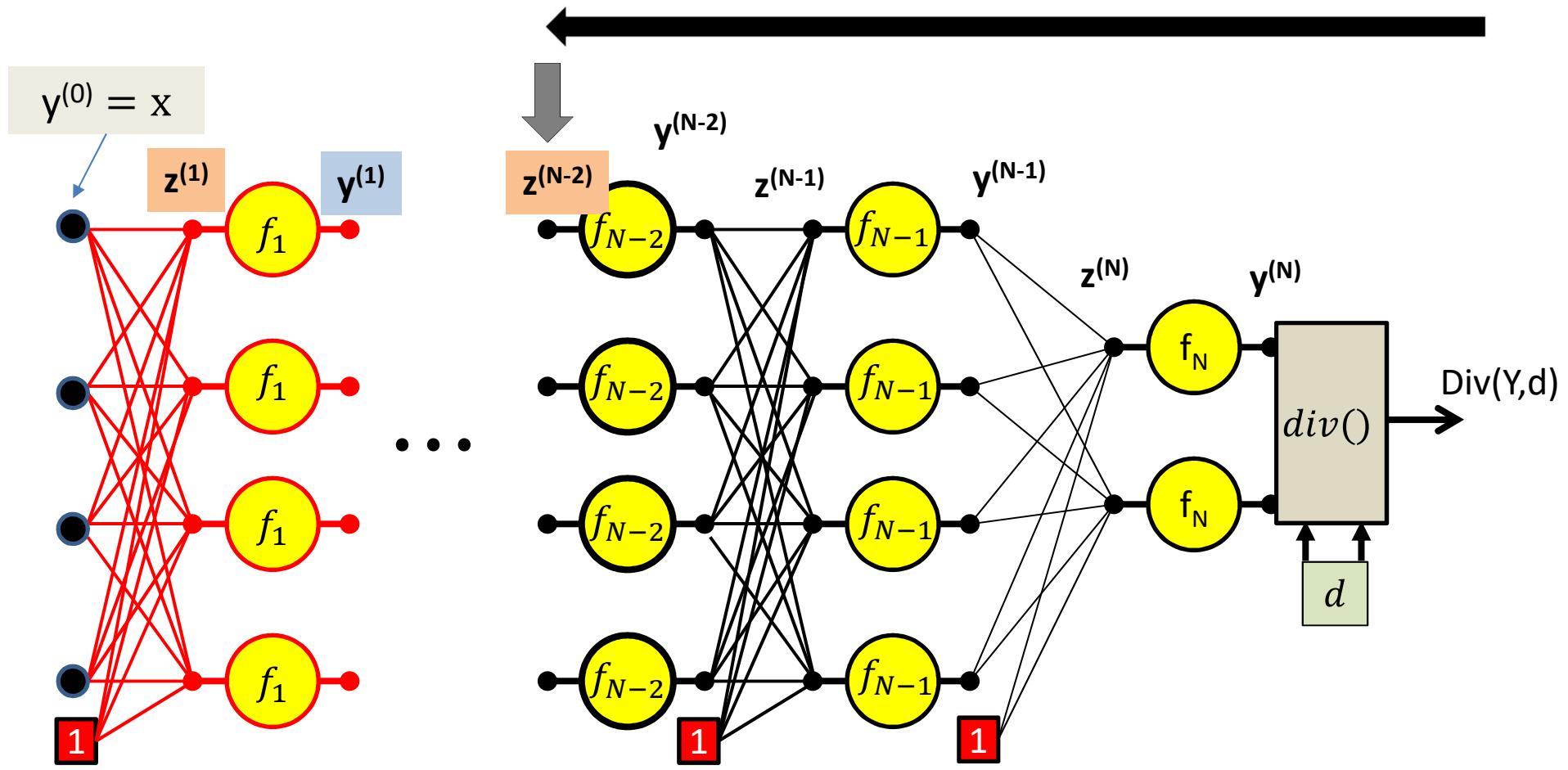
$$\frac{\partial \text{Div}}{\partial w_{ij}^{(N-1)}} = y_i^{(N-2)} \frac{\partial \text{Div}}{\partial z_j^{(N-1)}}$$

For the bias term  $y_0^{(N-2)} = 1$



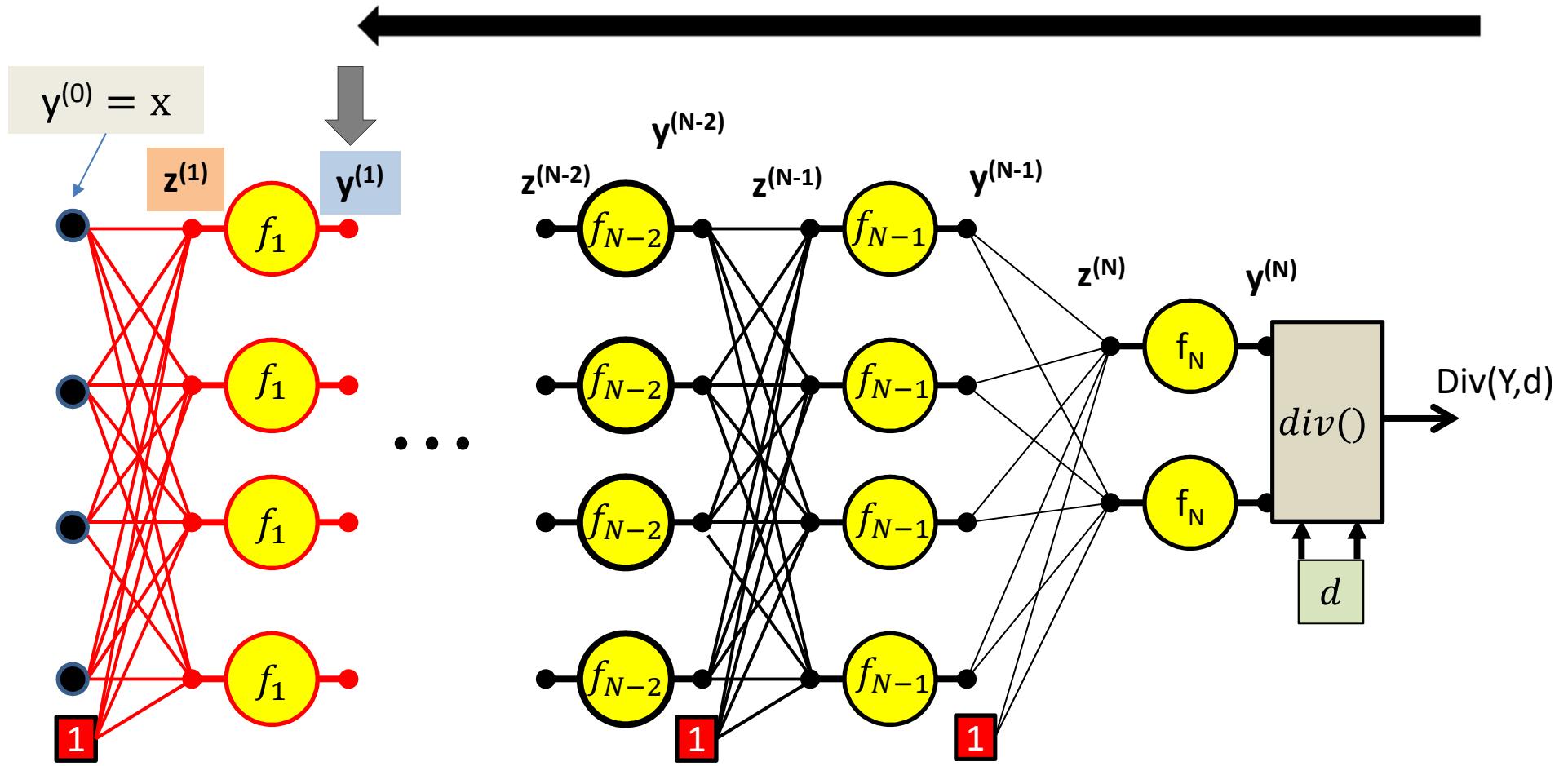
We continue our way backwards in the order shown

$$\frac{\partial Div}{\partial y_i^{(N-2)}} = \sum_j w_{ij}^{(N-1)} \frac{\partial Div}{\partial z_j^{(N-1)}}$$



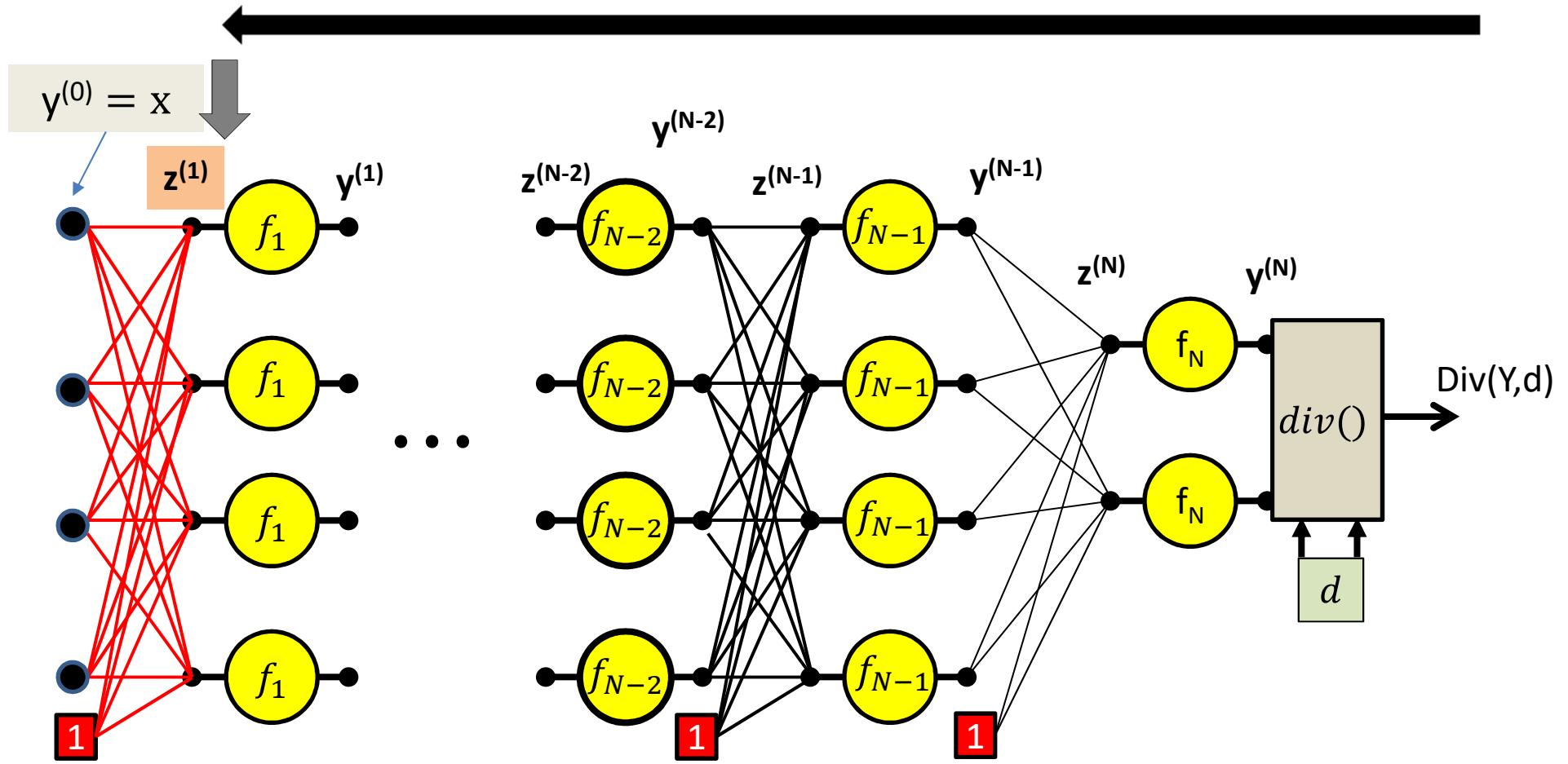
We continue our way backwards in the order shown

$$\frac{\partial Div}{\partial z_i^{(N-2)}} = f'_{N-2}(z_i^{(N-2)}) \frac{\partial Div}{\partial y_i^{(N-2)}}$$



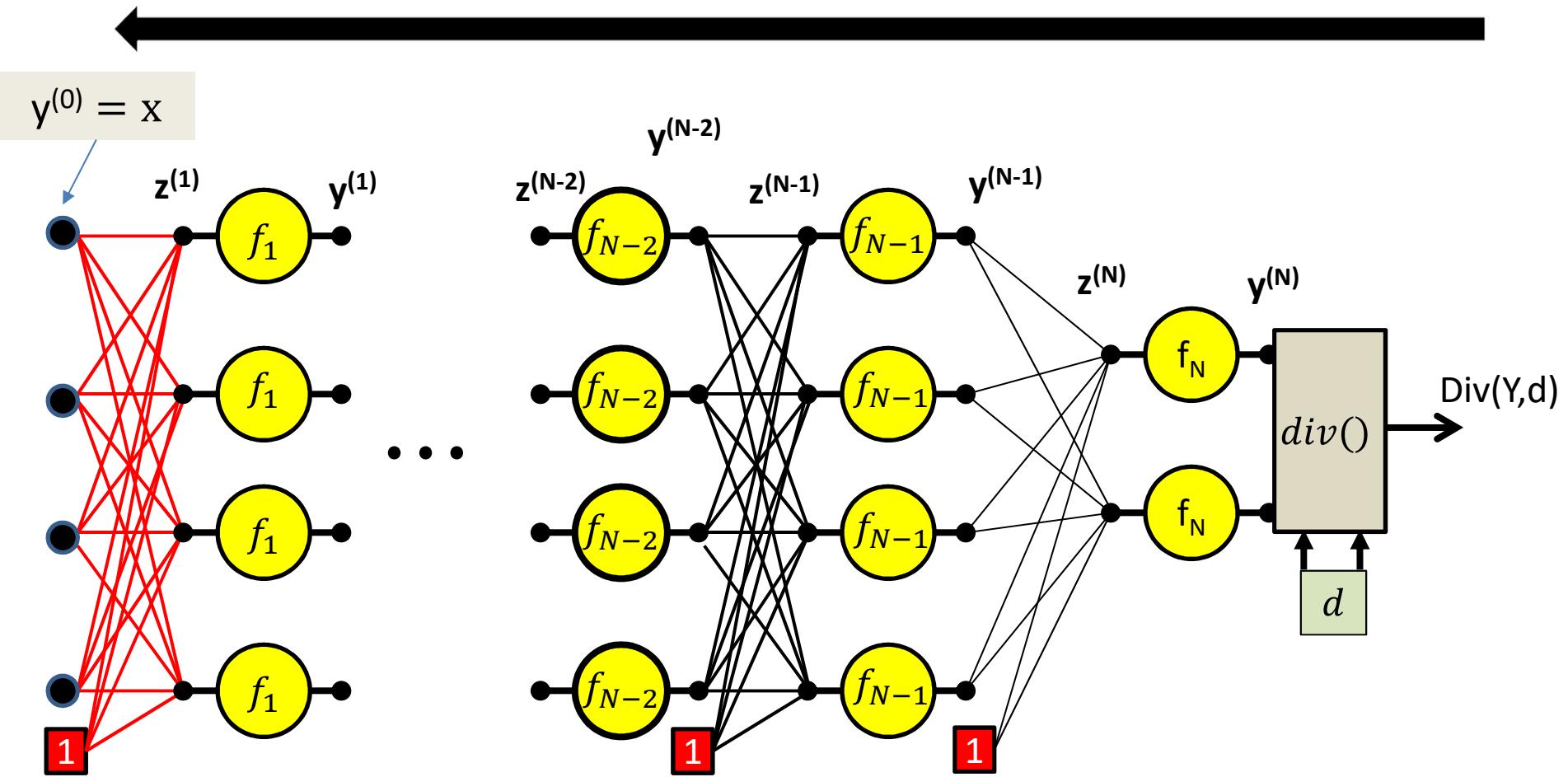
We continue our way backwards in the order shown

$$\frac{\partial Div}{\partial y_1^{(1)}} = \sum_j w_{ij}^{(2)} \frac{\partial Div}{\partial z_j^{(2)}}$$



We continue our way backwards in the order shown

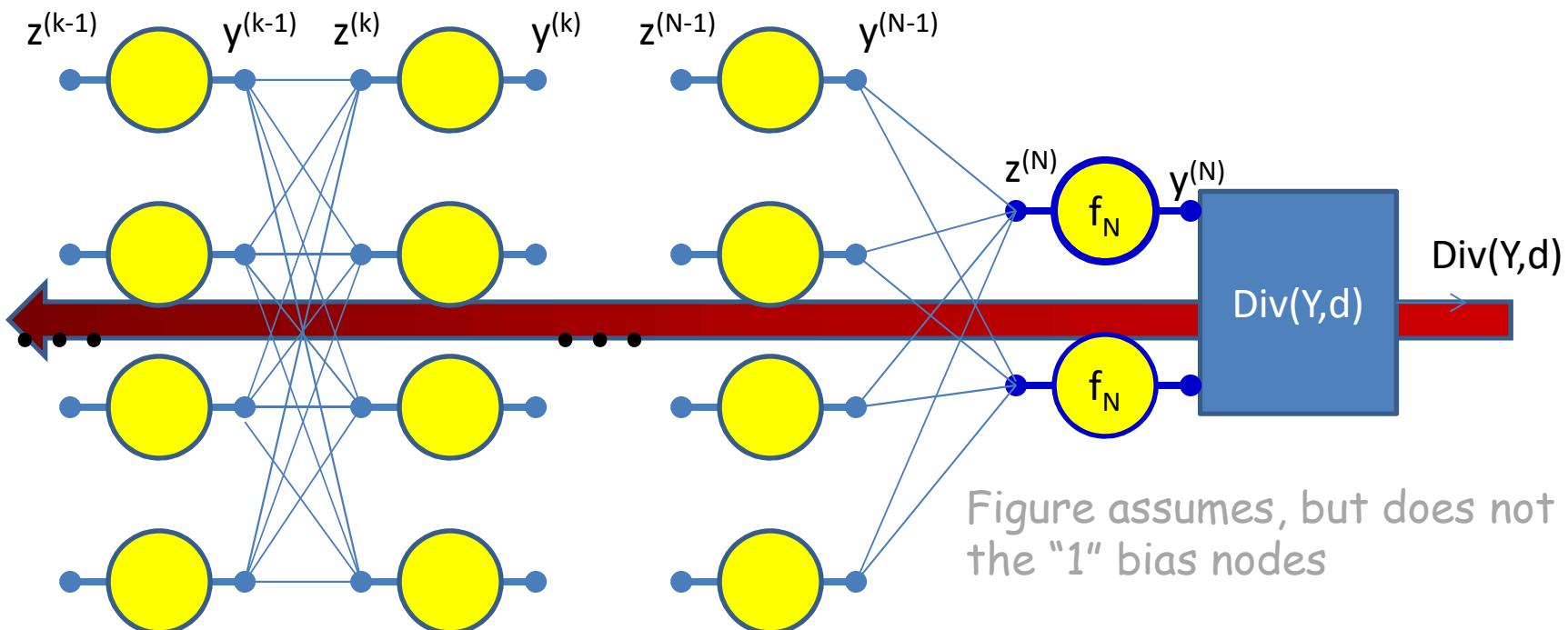
$$\frac{\partial Div}{\partial z_i^{(1)}} = f'_1(z_i^{(1)}) \frac{\partial Div}{\partial y_i^{(1)}}$$



We continue our way backwards in the order shown

$$\frac{\partial Div}{\partial w_{ij}^{(1)}} = y_i^{(0)} \frac{\partial Div}{\partial z_j^{(1)}}$$

# Gradients: Backward Computation



Initialize: Gradient  
w.r.t network output

$$\frac{\partial \text{Div}}{\partial y_i} = \frac{\partial \text{Div}(Y, d)}{\partial y_i^{(N)}}$$

$$\frac{\partial \text{Div}}{\partial z_i^{(N)}} = f'_k(z_i^{(N)}) \frac{\partial \text{Div}}{\partial y_i^{(N)}}$$

For  $k = N - 1..0$   
For  $i = 1: \text{layer width}$

$$\frac{\partial \text{Div}}{\partial y_i^{(k)}} = \sum_j w_{ij}^{(k+1)} \frac{\partial \text{Div}}{\partial z_j^{(k+1)}}$$

$$\frac{\partial \text{Div}}{\partial z_i^{(k)}} = f'_k(z_i^{(k)}) \frac{\partial \text{Div}}{\partial y_i^{(k)}}$$

$$\forall j \frac{\partial \text{Div}}{\partial w_{ij}^{(k+1)}} = y_i^{(k)} \frac{\partial \text{Div}}{\partial z_j^{(k+1)}}$$

# Backward Pass

- Output layer ( $N$ ) :
  - For  $i = 1 \dots D_N$ 
    - $\frac{\partial Div}{\partial y_i} = \frac{\partial Div(Y, d)}{\partial y_i^{(N)}}$
    - $\frac{\partial Div}{\partial z_i^{(N)}} = \frac{\partial Div}{\partial y_i^{(N)}} \frac{\partial y_i^{(N)}}{\partial z_i^{(N)}}$
- For layer  $k = N - 1$  down to 0
  - For  $i = 1 \dots D_k$ 
    - $\frac{\partial Div}{\partial y_i^{(k)}} = \sum_j w_{ij}^{(k+1)} \frac{\partial Div}{\partial z_j^{(k+1)}}$
    - $\frac{\partial Div}{\partial z_i^{(k)}} = \frac{\partial Di}{\partial y_i^{(k)}} f'_k(z_i^{(k)})$
    - $\frac{\partial Div}{\partial w_{ji}^{(k+1)}} = y_j^{(k)} \frac{\partial Div}{\partial z_i^{(k+1)}} \text{ for } j = 1 \dots D_{k+1}$

# Backward Pass

- Output layer ( $N$ ) :

- For  $i = 1 \dots D_N$

- $\frac{\partial Div}{\partial y_i} = \frac{\partial Div(Y, d)}{\partial y_i^{(N)}}$

- $\frac{\partial Div}{\partial z_i^{(N)}} = \frac{\partial Div}{\partial y_i^{(N)}} \frac{\partial y_i^{(N)}}{\partial z_i^{(N)}}$

Called “**Backpropagation**” because the derivative of the loss is propagated “backwards” through the network

- For layer  $k = N - 1$  down to 0

- For  $i = 1 \dots D_k$

- $\frac{\partial Div}{\partial y_i^{(k)}} = \sum_j w_{ij}^{(k+1)} \frac{\partial Div}{\partial z_j^{(k+1)}}$

Backward weighted combination of next layer

- $\frac{\partial Div}{\partial z_i^{(k)}} = \frac{\partial Di}{\partial y_i^{(k)}} f'_k(z_i^{(k)})$

Backward equivalent of activation

- $\frac{\partial Div}{\partial w_{ji}^{(k+1)}} = y_j^{(k)} \frac{\partial Div}{\partial z_i^{(k+1)}}$  for  $j = 1 \dots D_{k+1}$

Using notation  $\dot{y} = \frac{\partial \text{Div}(Y, d)}{\partial y}$  etc (overdot represents derivative of  $\text{Div}$  w.r.t variable)

- Output layer ( $N$ ) :

- For  $i = 1 \dots D_N$

- $\dot{y}_i = \frac{\partial \text{Div}(Y, d)}{\partial y_i^{(N)}}$

- $\dot{z}_i^{(N)} = \dot{y}_i \frac{\partial y_i^{(N)}}{\partial z_i^{(N)}}$

Called "Backpropagation" because the derivative of the loss is propagated "backwards" through the network

Very analogous to the forward pass:

- For layer  $k = N - 1$  down to 0

- For  $i = 1 \dots D_k$

- $\dot{y}_i^{(k)} = \sum_j w_{ij}^{(k+1)} \dot{z}_j^{(k+1)}$

Backward weighted combination of next layer

- $\dot{z}_i^{(k)} = \dot{y}_i^{(i)} f'_k(z_i^{(k)})$

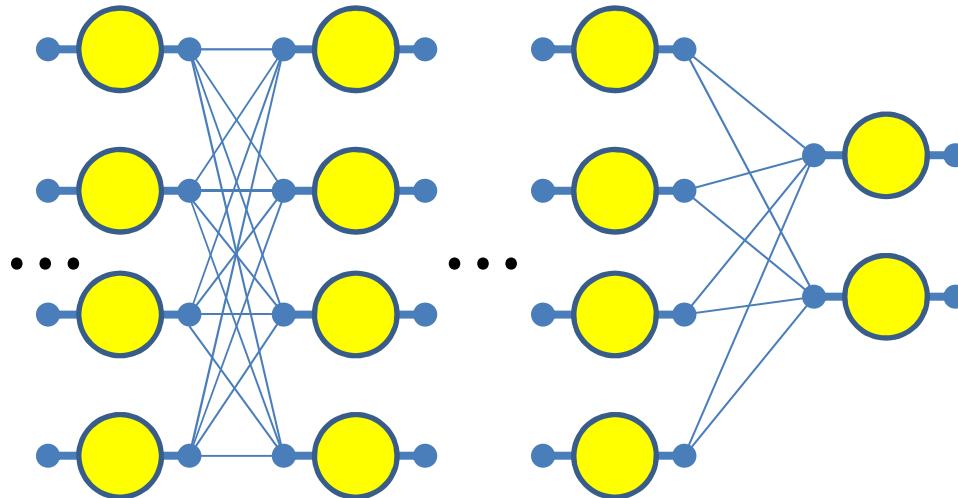
Backward equivalent of activation

- $\frac{\partial \text{Div}}{\partial w_{ji}^{(k+1)}} = y_j^{(k)} \dot{z}_i^{(k+1)}$  for  $j = 1 \dots D_{k+1}$

# For comparison: the forward pass again

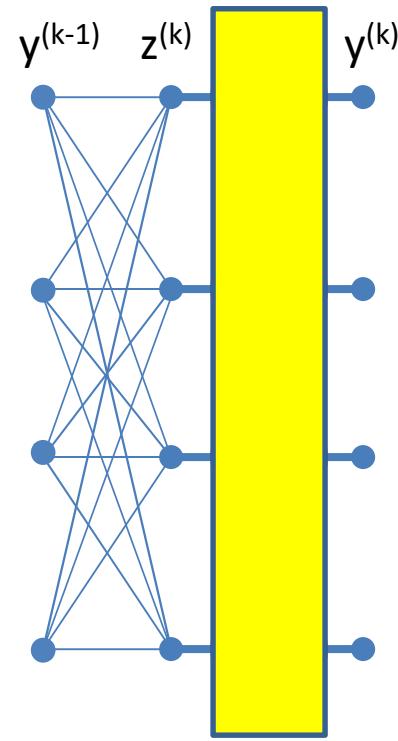
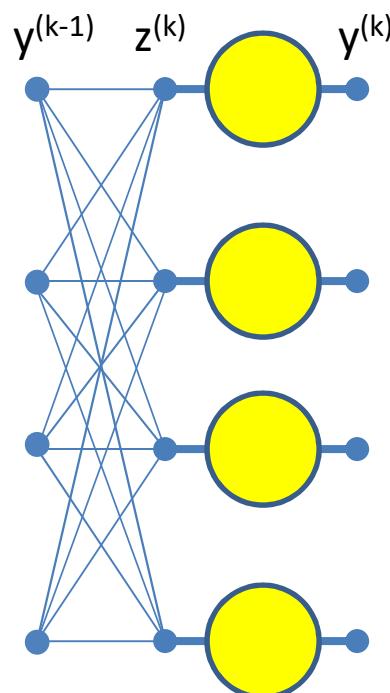
- Input:  $D$  dimensional vector  $\mathbf{x} = [x_j, j = 1 \dots D]$
- Set:
  - $D_0 = D$ , is the width of the 0<sup>th</sup> (input) layer
  - $y_j^{(0)} = x_j, j = 1 \dots D; y_0^{(k=1\dots N)} = x_0 = 1$
- For layer  $k = 1 \dots N$ 
  - For  $j = 1 \dots D_k$ 
    - $z_j^{(k)} = \sum_{i=0}^{N_k} w_{i,j}^{(k)} y_i^{(k-1)}$
    - $y_j^{(k)} = f_k(z_j^{(k)})$
- Output:
  - $Y = y_j^{(N)}, j = 1..D_N$

# Special cases



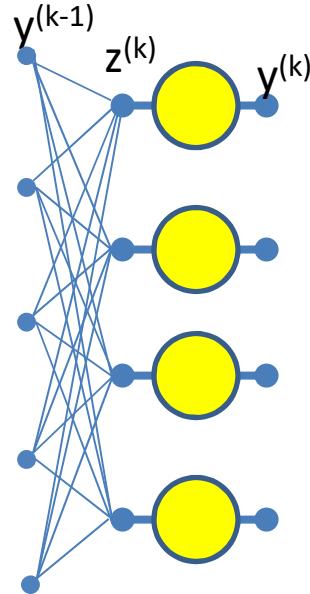
- Have assumed so far that
  1. The computation of the output of one neuron does not directly affect computation of other neurons in the same (or previous) layers
  2. Outputs of neurons only combine through weighted addition
  3. Activations are actually differentiable
    - All of these conditions are frequently not applicable
- Will not dwell on the topic in class, but explained in slides
  - Will appear in quiz. Please read the slides

# Special Case 1. Vector activations



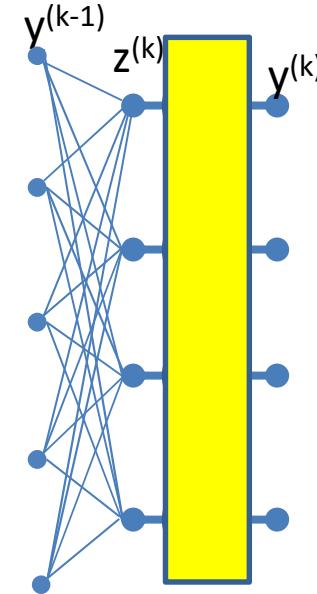
- Vector activations: all outputs are functions of all inputs

# Special Case 1. Vector activations



Scalar activation: Modifying a  $z_i$  only changes corresponding  $y_i$

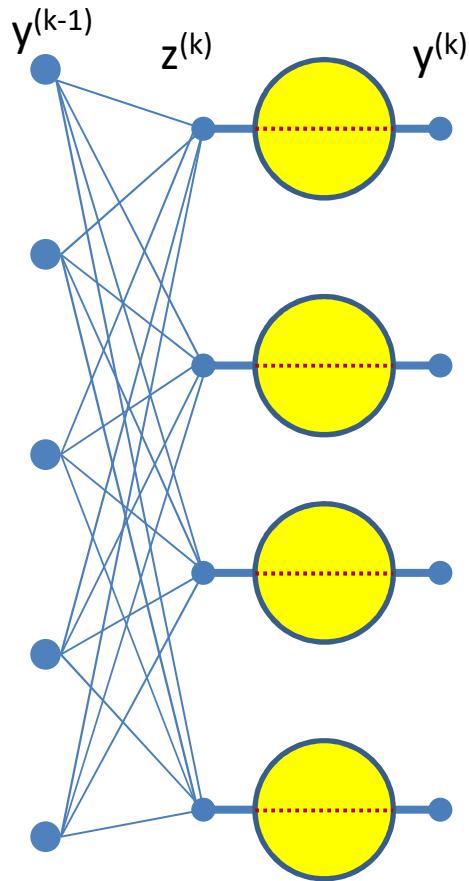
$$y_i^{(k)} = f(z_i^{(k)})$$



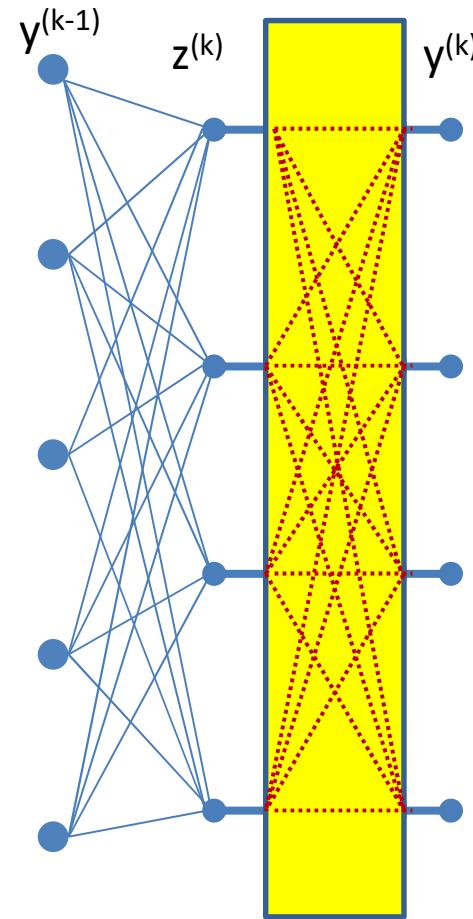
Vector activation: Modifying a  $z_i$  potentially changes all,  $y_1 \dots y_M$

$$\begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \vdots \\ y_M^{(k)} \end{bmatrix} = f \left( \begin{bmatrix} z_1^{(k)} \\ z_2^{(k)} \\ \vdots \\ z_D^{(k)} \end{bmatrix} \right)$$

# “Influence” diagram

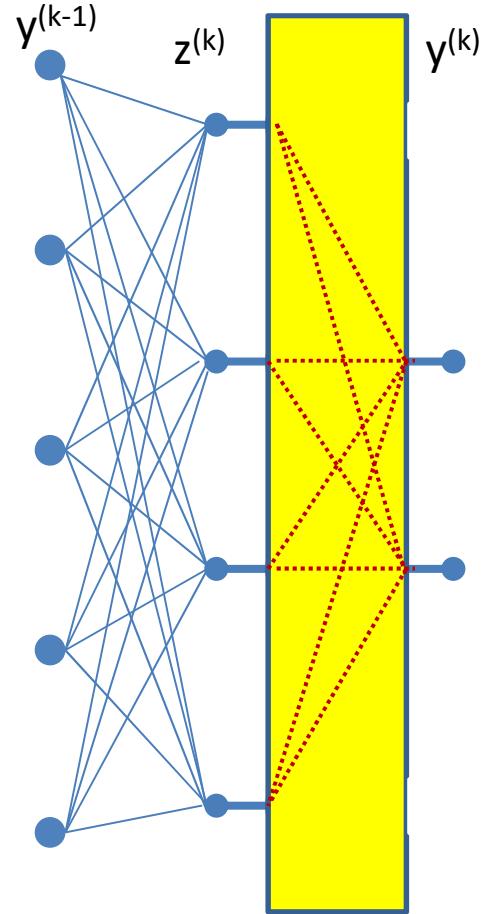
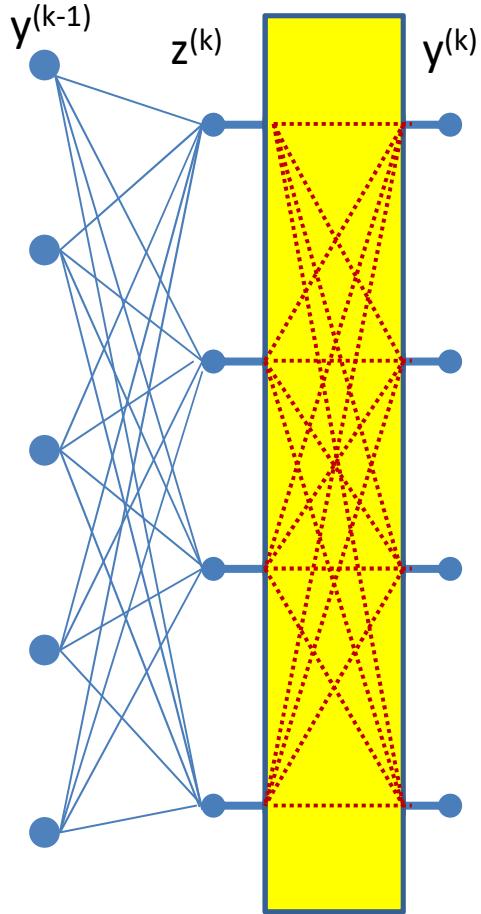


Scalar activation: Each  $z_i$  influences *one*  $y_i$



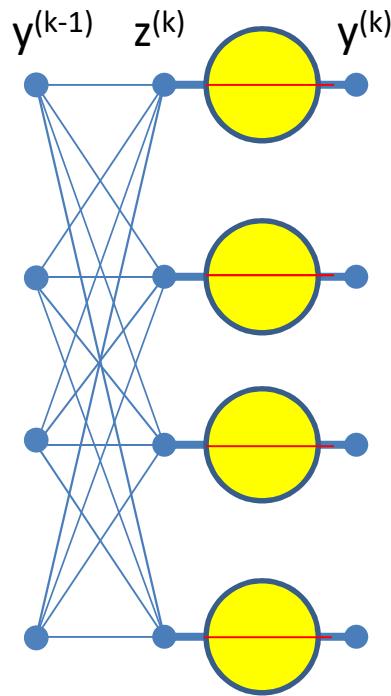
Vector activation: Each  $z_i$  influences all,  $y_1 \dots y_M$

# The number of outputs



- Note: The number of outputs ( $y^{(k)}$ ) need not be the same as the number of inputs ( $z^{(k)}$ )
  - May be more or fewer

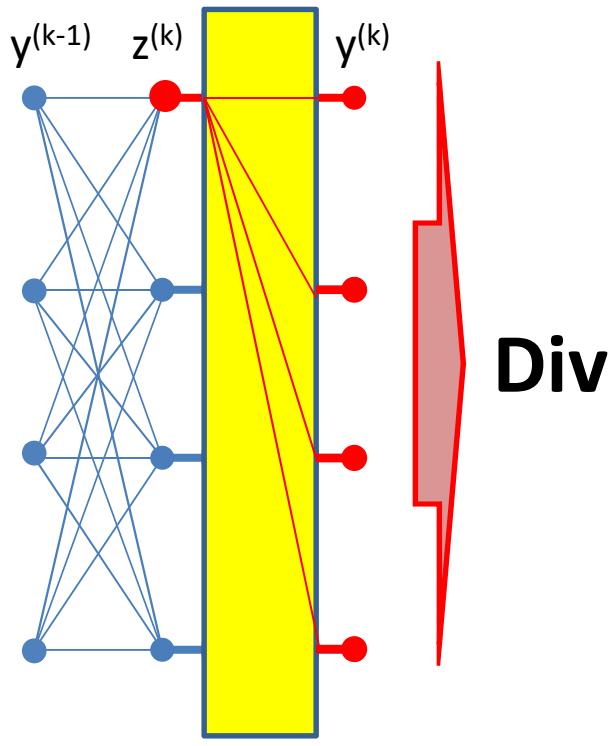
# Scalar Activation: Derivative rule



$$\frac{\partial \text{Div}}{\partial z_i^{(k)}} = \frac{\partial \text{Div}}{\partial y_i^{(k)}} \frac{dy_i^{(k)}}{dz_i^{(k)}}$$

- In the case of *scalar* activation functions, the derivative of the error w.r.t to the input to the unit is a simple product of derivatives

# Derivatives of vector activation



$$\frac{\partial \text{Div}}{\partial z_i^{(k)}} = \sum_j \frac{\partial \text{Div}}{\partial y_j^{(k)}} \frac{\partial y_j^{(k)}}{\partial z_i^{(k)}}$$

Note: derivatives of scalar activations are just a special case of vector activations:

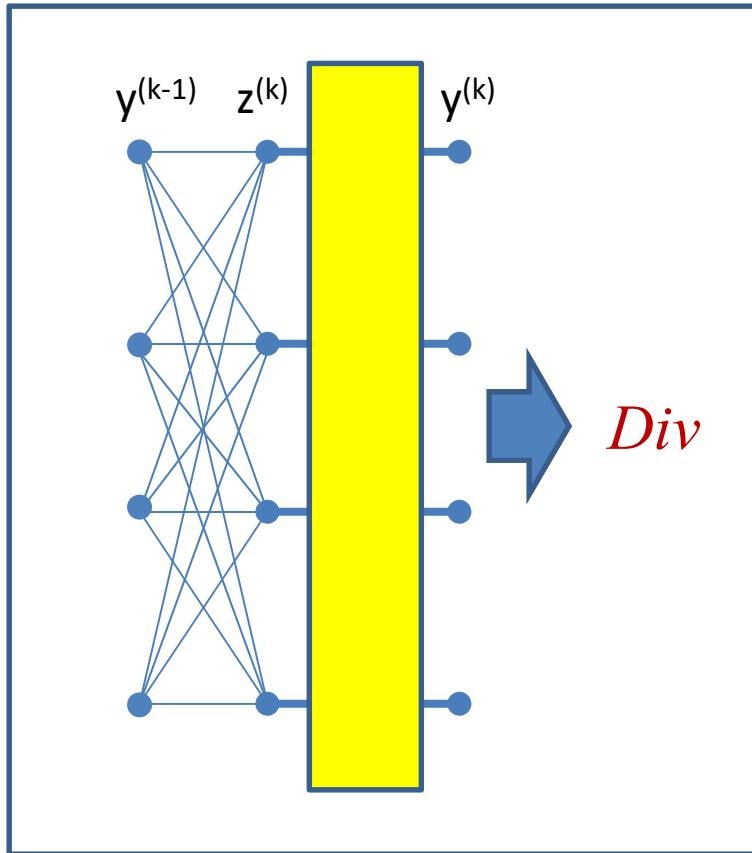
$$\frac{\partial y_j^{(k)}}{\partial z_i^{(k)}} = 0 \text{ for } i \neq j$$

- For *vector* activations the derivative of the error w.r.t. to any input is a sum of partial derivatives
  - Regardless of the number of outputs  $y_j^{(k)}$

# Special cases

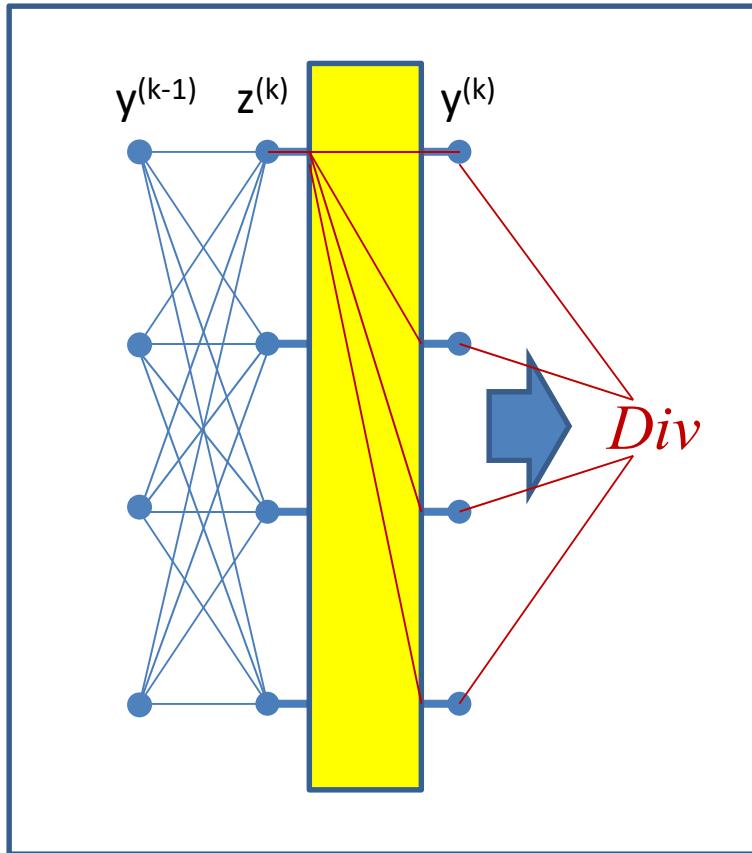
- Examples of vector activations and other special cases on slides
  - Please look up
  - Will appear in quiz!

# Example Vector Activation: Softmax



$$y_i^{(k)} = \frac{\exp(z_i^{(k)})}{\sum_j \exp(z_j^{(k)})}$$

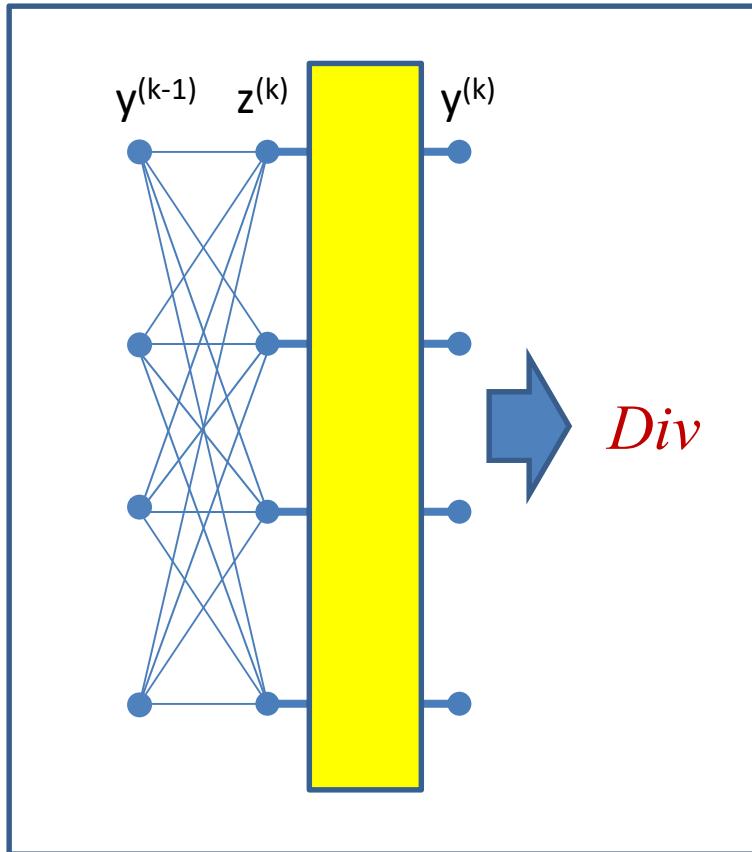
# Example Vector Activation: Softmax



$$y_i^{(k)} = \frac{\exp(z_i^{(k)})}{\sum_j \exp(z_j^{(k)})}$$

$$\frac{\partial Div}{\partial z_i^{(k)}} = \sum_j \frac{\partial Div}{\partial y_j^{(k)}} \frac{\partial y_j^{(k)}}{\partial z_i^{(k)}}$$

# Example Vector Activation: Softmax

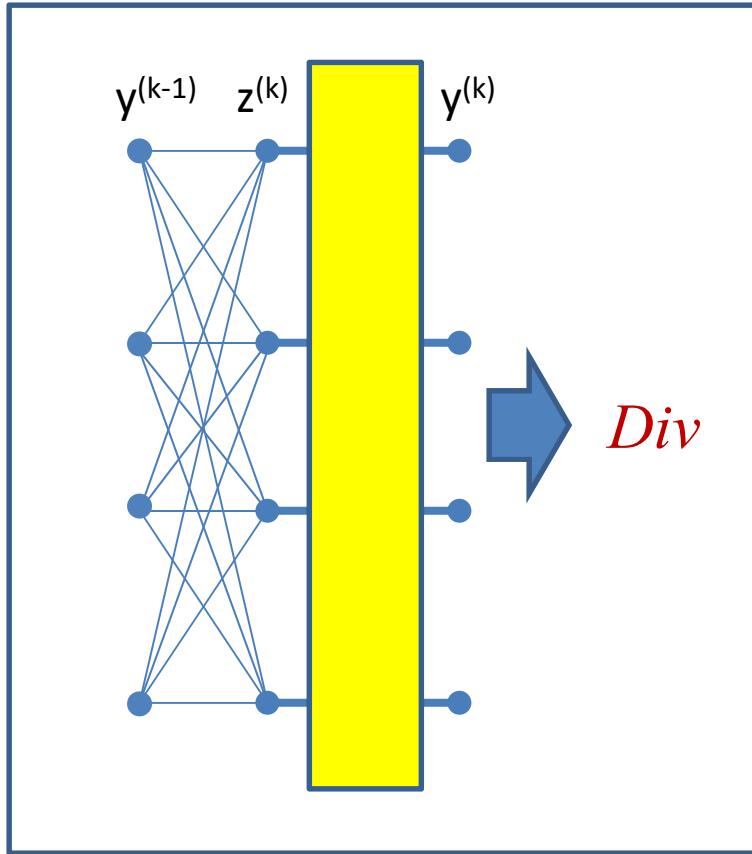


$$y_i^{(k)} = \frac{\exp(z_i^{(k)})}{\sum_j \exp(z_j^{(k)})}$$

$$\frac{\partial Div}{\partial z_i^{(k)}} = \sum_j \frac{\partial Div}{\partial y_j^{(k)}} \frac{\partial y_j^{(k)}}{\partial z_i^{(k)}}$$

$$\frac{\partial y_j^{(k)}}{\partial z_i^{(k)}} = \begin{cases} y_i^{(k)} (1 - y_i^{(k)}) & \text{if } i = j \\ -y_i^{(k)} y_j^{(k)} & \text{if } i \neq j \end{cases}$$

# Example Vector Activation: Softmax



$$y_i^{(k)} = \frac{\exp(z_i^{(k)})}{\sum_j \exp(z_j^{(k)})}$$

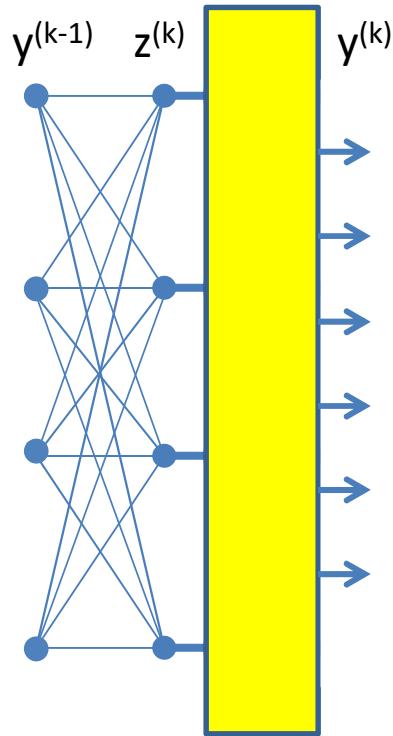
$$\frac{\partial \text{Div}}{\partial z_i^{(k)}} = \sum_j \frac{\partial \text{Div}}{\partial y_j^{(k)}} \frac{\partial y_j^{(k)}}{\partial z_i^{(k)}}$$

$$\frac{\partial y_j^{(k)}}{\partial z_i^{(k)}} = \begin{cases} y_i^{(k)} (1 - y_i^{(k)}) & \text{if } i = j \\ -y_i^{(k)} y_j^{(k)} & \text{if } i \neq j \end{cases}$$

$$\frac{\partial \text{Div}}{\partial z_i^{(k)}} = \sum_j \frac{\partial \text{Div}}{\partial y_j^{(k)}} y_i^{(k)} (\delta_{ij} - y_j^{(k)})$$

- For future reference
- $\delta_{ij}$  is the Kronecker delta:  $\delta_{ij} = 1$  if  $i = j$ ,  $0$  if  $i \neq j$

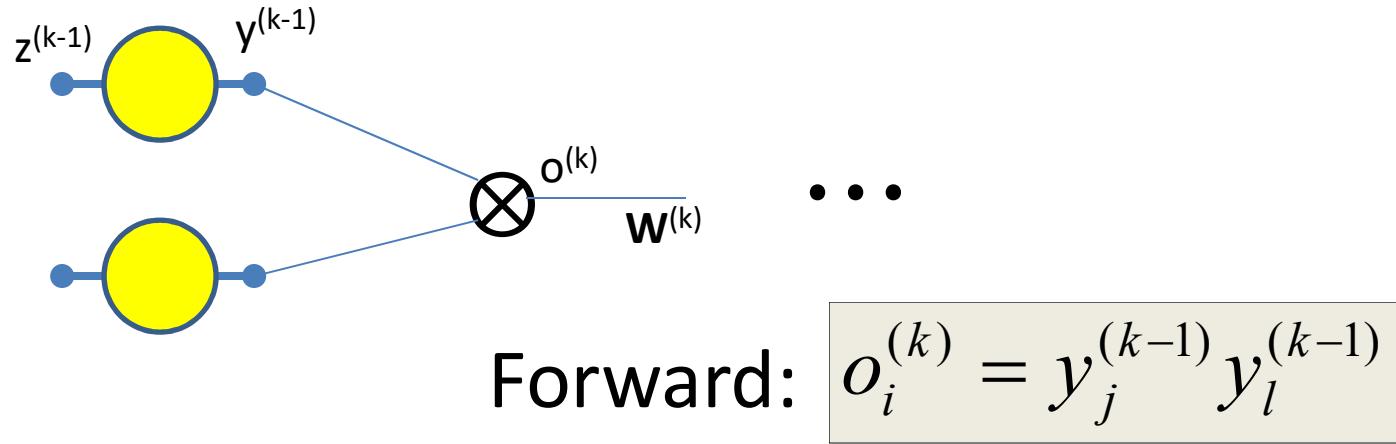
# Vector Activations



$$\begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \vdots \\ y_M^{(k)} \end{bmatrix} = f \left( \begin{bmatrix} z_1^{(k)} \\ z_2^{(k)} \\ \vdots \\ z_D^{(k)} \end{bmatrix} \right)$$

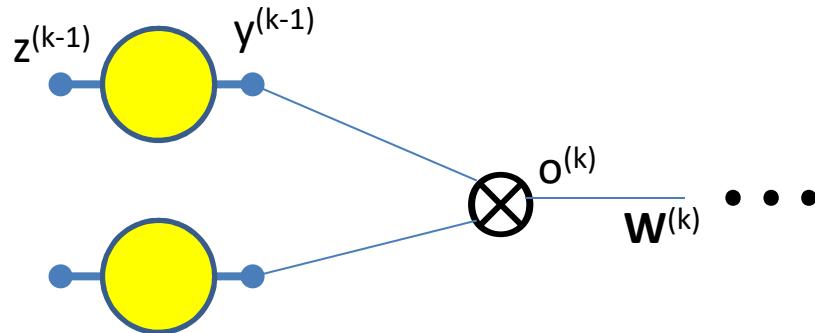
- In reality the vector combinations can be anything
  - E.g. linear combinations, polynomials, logistic (softmax), etc.

# Special Case 2: Multiplicative networks



- Some types of networks have *multiplicative* combination
  - In contrast to the *additive* combination we have seen so far
- Seen in networks such as LSTMs, GRUs, attention models, etc.

# Backpropagation: Multiplicative Networks



Forward:

$$o_i^{(k)} = y_j^{(k-1)} y_l^{(k-1)}$$

Backward:

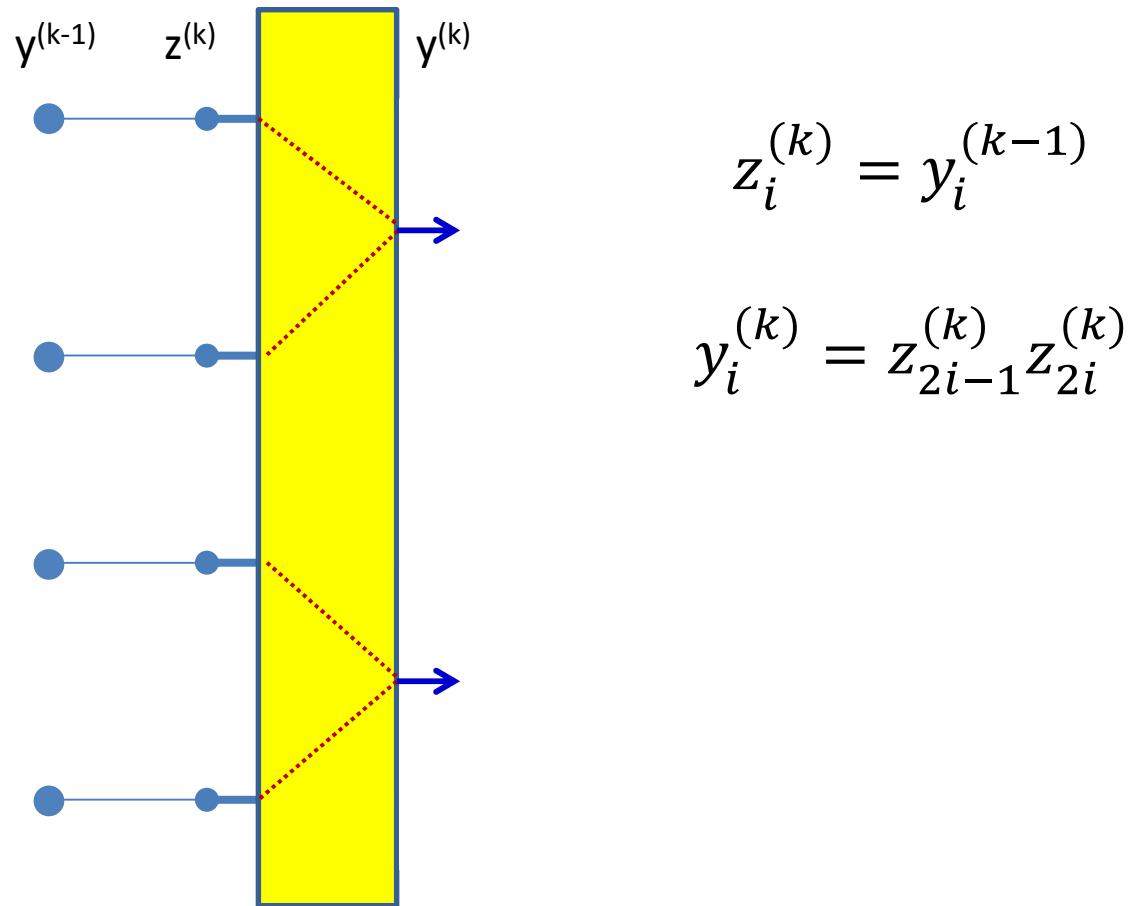
$$\frac{\partial \text{Div}}{\partial o_i^{(k)}} = \sum_j w_{ij}^{(k+1)} \frac{\partial \text{Div}}{\partial z_j^{(k+1)}}$$

$$\frac{\partial \text{Div}}{\partial y_j^{(k-1)}} = \frac{\partial o_i^{(k)}}{\partial y_j^{(k-1)}} \frac{\partial \text{Div}}{\partial o_i^{(k)}} = y_l^{(k-1)} \frac{\partial \text{Div}}{\partial o_i^{(k)}}$$

$$\frac{\partial \text{Div}}{\partial y_l^{(k-1)}} = y_j^{(k-1)} \frac{\partial \text{Div}}{\partial o_i^{(k)}}$$

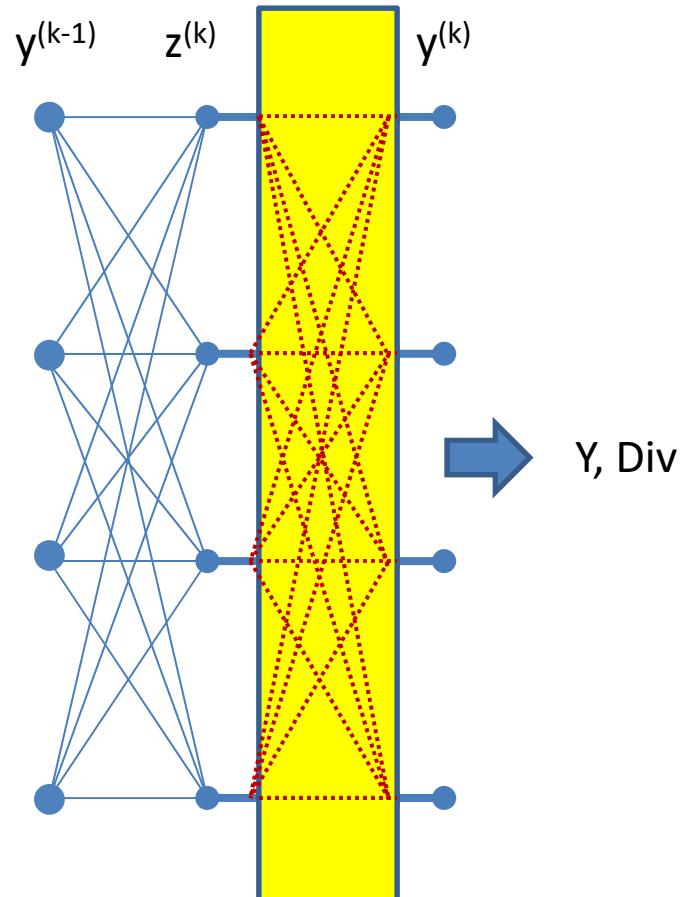
- Some types of networks have *multiplicative* combination

# Multiplicative combination as a case of vector activations



- A layer of multiplicative combination is a special case of vector activation

# Multiplicative combination: Can be viewed as a case of vector activations



$$z_i^{(k)} = \sum_j w_{ji}^{(k)} y_j^{(k-1)}$$

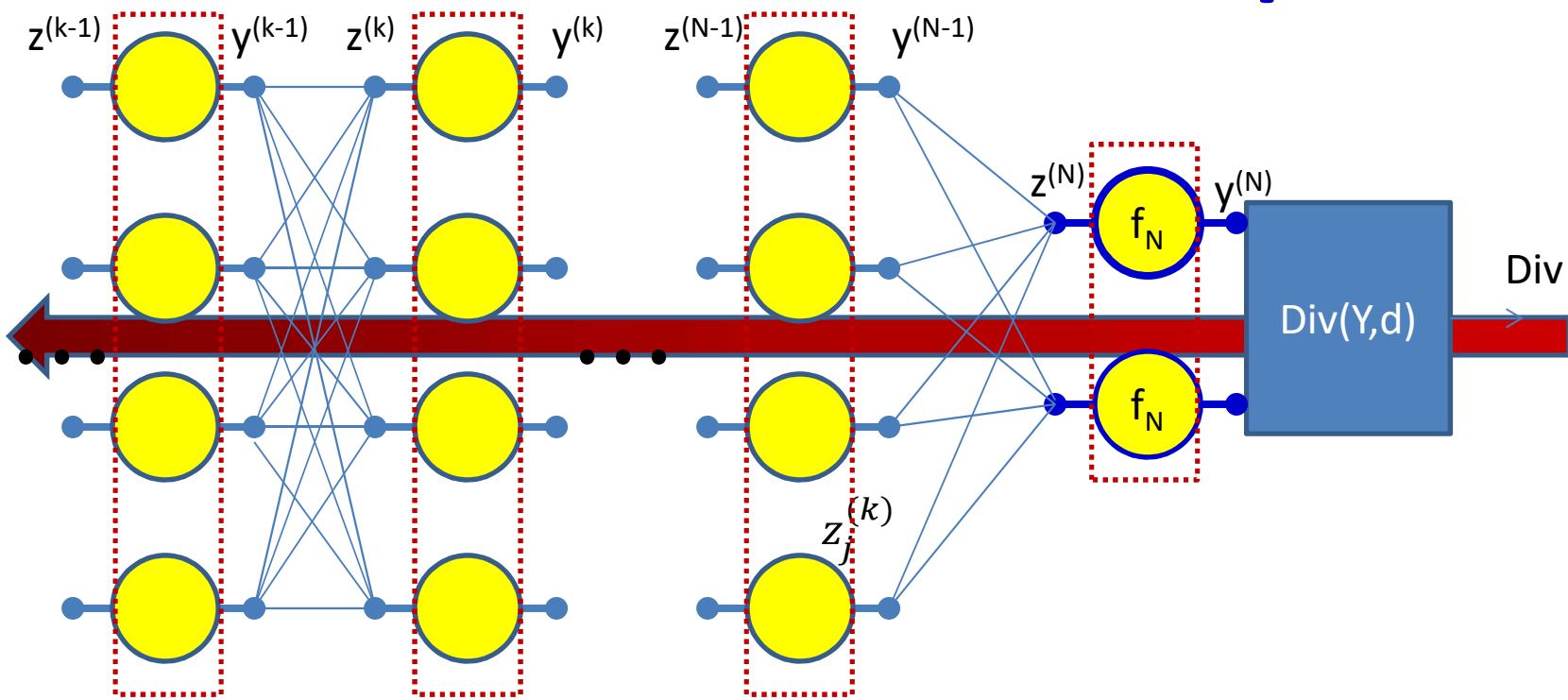
$$y_i^{(k)} = \prod_l (z_l^{(k)})^{\alpha_{li}^{(k)}}$$

$$\frac{\partial y_i^{(k)}}{\partial z_j^{(k)}} = \alpha_{ji}^{(k)} (z_j^{(k)})^{\alpha_{ji}^{(k)} - 1} \prod_{l \neq j} (z_l^{(k)})^{\alpha_{li}^{(k)}}$$

$$\frac{\partial Div}{\partial z_j^{(k)}} = \sum_i \frac{\partial Div}{\partial y_i^{(k)}} \frac{\partial y_i^{(k)}}{\partial z_j^{(k)}}$$

- A layer of multiplicative combination is a special case of vector activation

# Gradients: Backward Computation



For  $k = N \dots 1$

For  $i = 1 : \text{layer width}$

If layer has vector activation

$$\frac{\partial \text{Div}}{\partial z_i^{(k)}} = \sum_j \frac{\partial \text{Div}}{\partial y_j^{(k)}} \frac{\partial y_j^{(k)}}{\partial z_i^{(k)}}$$

Else if activation is scalar

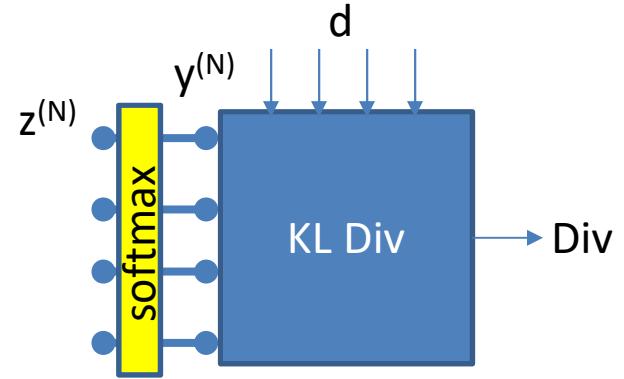
$$\frac{\partial \text{Div}}{\partial z_i^{(k)}} = \frac{\partial \text{Div}}{\partial y_i^{(k)}} \frac{\partial y_i^{(k)}}{\partial z_i^{(k)}}$$

$$\frac{\partial \text{Div}}{\partial y_i^{(k-1)}} = \sum_j w_{ij}^{(k)} \frac{\partial \text{Div}}{\partial z_j^{(k)}}$$

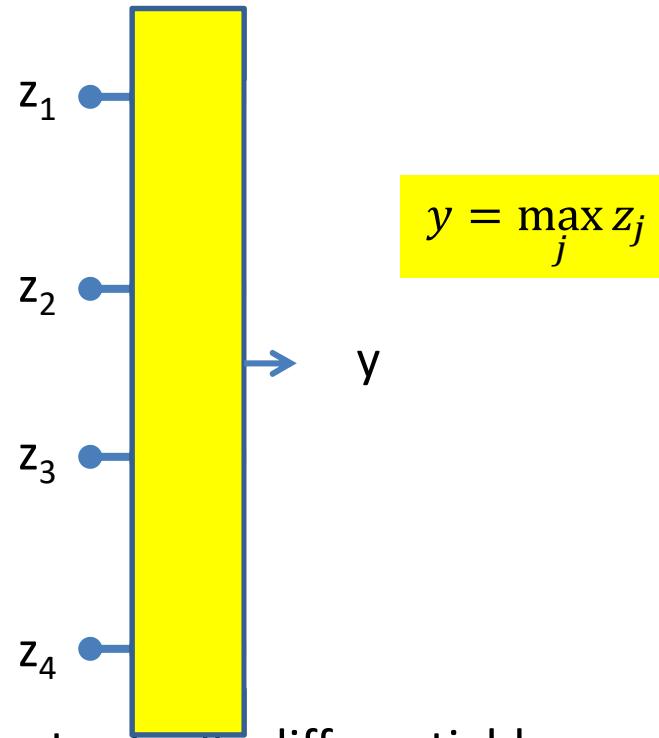
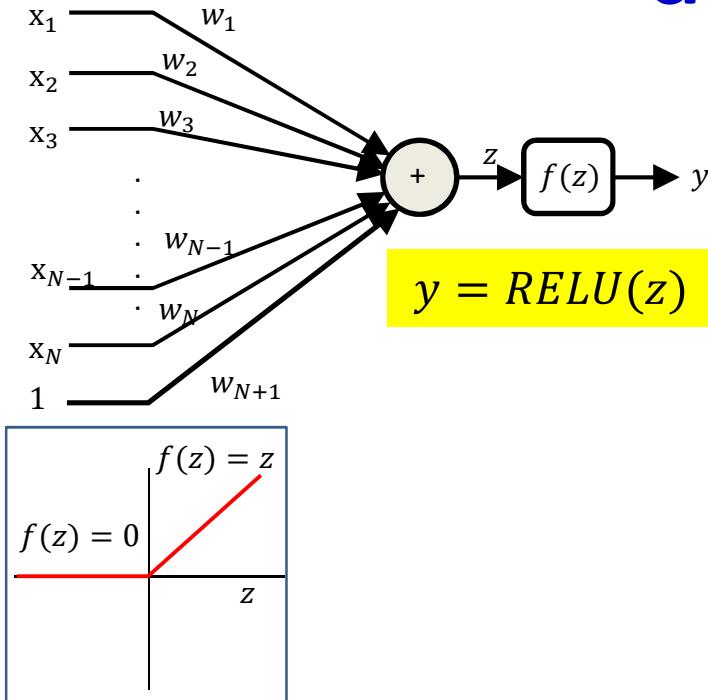
$$\frac{\partial \text{Div}}{\partial w_{ij}^{(k)}} = y_i^{(k-1)} \frac{\partial \text{Div}}{\partial z_j^{(k)}}$$

# Backward Pass for softmax output layer

- Output layer ( $N$ ) :
  - For  $i = 1 \dots D_N$ 
    - $\frac{\partial Di}{\partial y_i} = \frac{\partial Div(Y, d)}{\partial y_i^{(N)}}$
    - $\frac{\partial Div}{\partial z_i^{(N)}} = \sum_j \frac{\partial Div(Y, d)}{\partial y_j^{(N)}} y_i^{(N)} (\delta_{ij} - y_j^{(N)})$
- For layer  $k = N - 1$  down to 0
  - For  $i = 1 \dots D_k$ 
    - $\frac{\partial Div}{\partial y_i^{(k)}} = \sum_j w_{ij}^{(k+1)} \frac{\partial Div}{\partial z_j^{(k+1)}}$
    - $\frac{\partial Div}{\partial z_i^{(k)}} = f'_k(z_i^{(k)}) \frac{\partial Div}{\partial y_i^{(k)}}$
    - $\frac{\partial Div}{\partial w_{ji}^{(k+1)}} = y_j^{(k)} \frac{\partial Div}{\partial z_i^{(k+1)}} \text{ for } j = 1 \dots D_{k+1}$

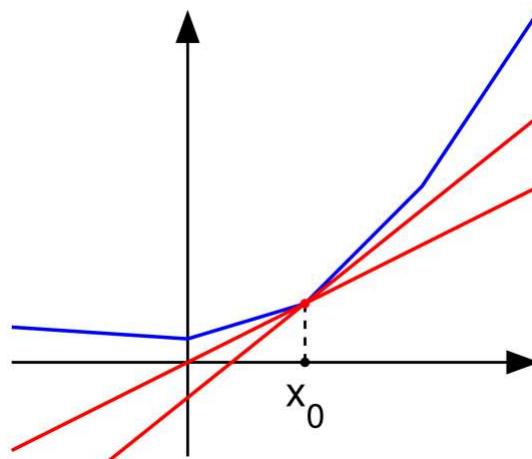


# Special Case 3: Non-differentiable activations



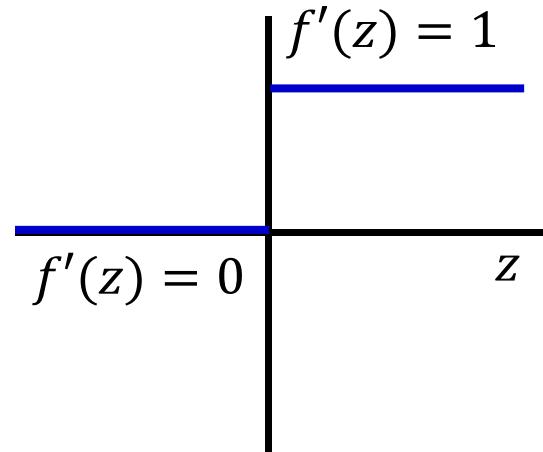
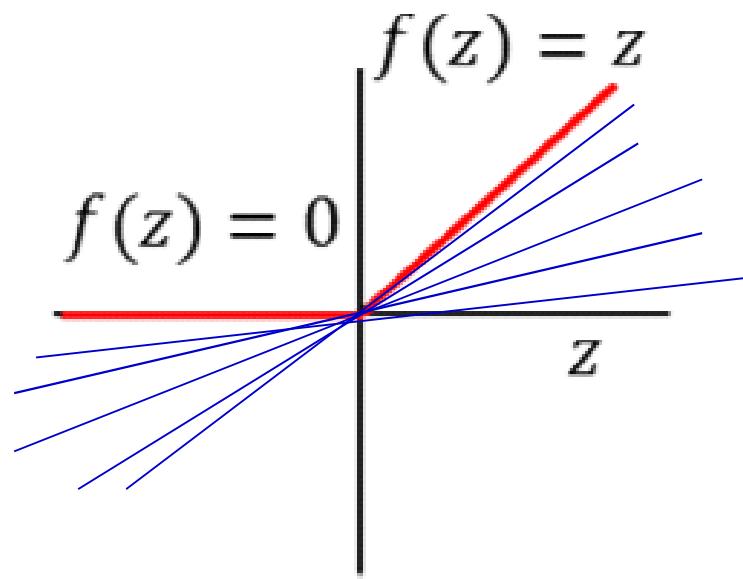
- Activation functions are sometimes not actually differentiable
  - E.g. The RELU (Rectified Linear Unit)
    - And its variants: leaky RELU, randomized leaky RELU
  - E.g. The “max” function
- Must use “subgradients” where available
  - Or “secants”

# The subgradient



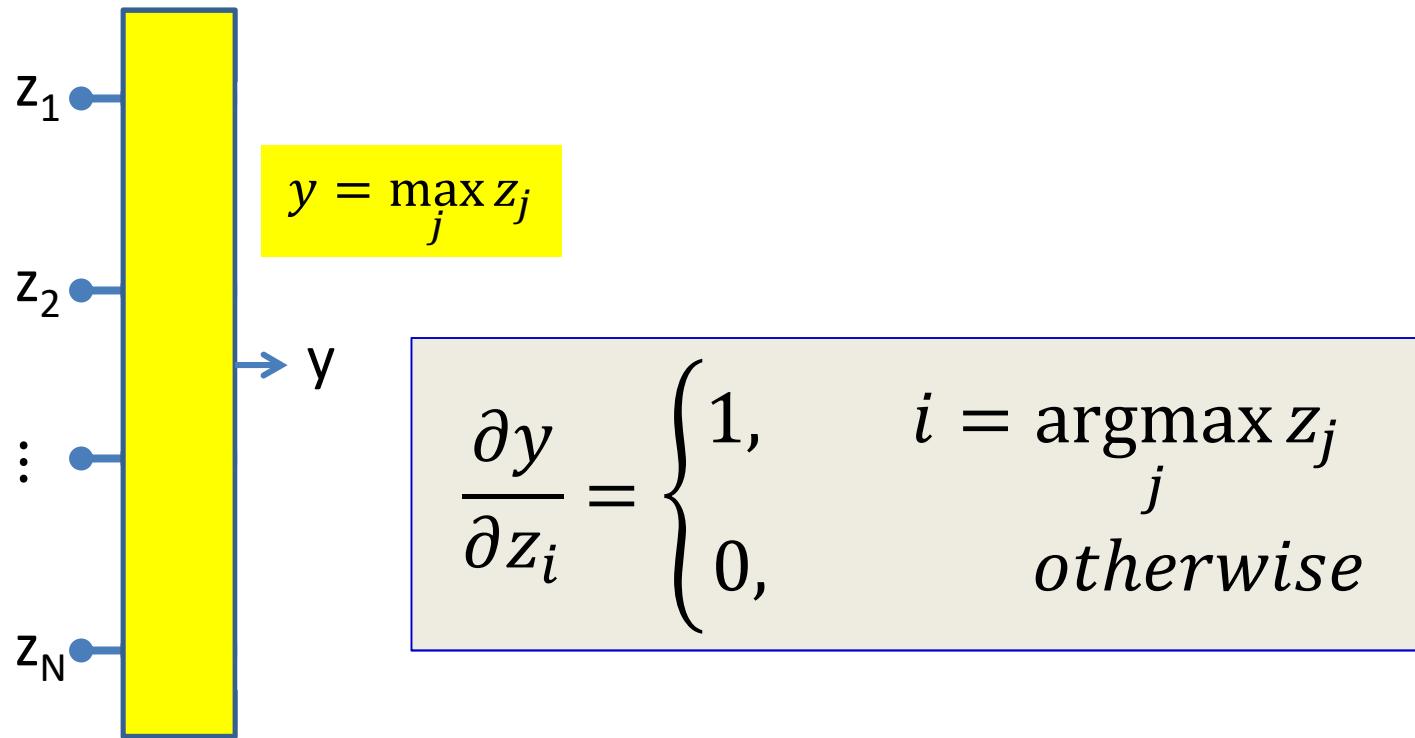
- A subgradient of a function  $f(x)$  at a point  $x_0$  is any vector  $v$  such that
$$(f(x) - f(x_0)) \geq v^T(x - x_0)$$
  - Any direction such that moving in that direction increases the function
- Guaranteed to exist only for convex functions
  - “bowl” shaped functions
  - For non-convex functions, the equivalent concept is a “quasi-secant”
- The subgradient is a direction in which the function is guaranteed to increase
- If the function is differentiable at  $x_0$ , the subgradient is the gradient
  - The gradient is not always the subgradient though

# Subgradients and the RELU



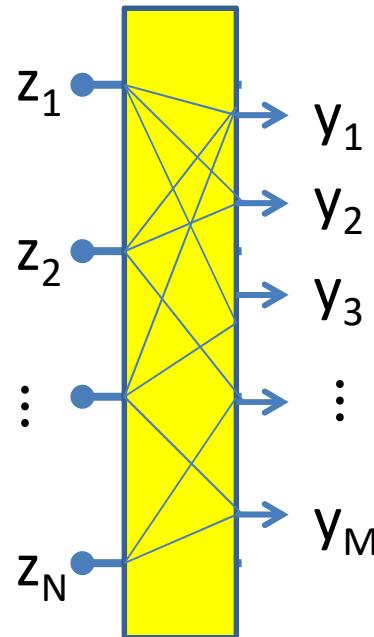
- Can use any subgradient
  - At the differentiable points on the curve, this is the same as the gradient
  - Typically, will use the equation given

# Subgradients and the Max



- Vector equivalent of subgradient
  - 1 w.r.t. the largest incoming input
    - Incremental changes in this input will change the output
  - 0 for the rest
    - Incremental changes to these inputs will not change the output

# Subgradients and the Max



$$y_i = \operatorname{argmax}_{l \in \mathcal{S}_j} z_l$$

$$\frac{\partial y_j}{\partial z_i} = \begin{cases} 1, & i = \operatorname{argmax}_{l \in \mathcal{S}_j} z_l \\ 0, & \text{otherwise} \end{cases}$$

- Multiple outputs, each selecting the max of a different subset of inputs
  - Will be seen in convolutional networks
- Gradient for any output:
  - 1 for the specific component that is maximum in corresponding input subset
  - 0 otherwise

# Backward Pass: Recap

- Output layer ( $N$ ) :

– For  $i = 1 \dots D_N$

- $\frac{\partial \text{Div}}{\partial Y_i} = \frac{\partial \text{Div}(Y, d)}{\partial y_i^{(N)}}$

- $\frac{\partial \text{Div}}{\partial z_i^{(N)}} = \frac{\partial \text{Div}}{\partial y_i^{(N)}} \frac{\partial y_i^{(N)}}{\partial z_i^{(N)}} \quad OR \quad \sum_j \frac{\partial \text{Div}}{\partial y_j^{(N)}} \frac{\partial y_j^{(N)}}{\partial z_i^{(N)}} \quad (\text{vector activation})$

- For layer  $k = N - 1$  down to 0

– For  $i = 1 \dots D_k$

- $\frac{\partial \text{Div}}{\partial y_i^{(k)}} = \sum_j w_{ij}^{(k+1)} \frac{\partial \text{Div}}{\partial z_j^{(k+1)}}$

- $\frac{\partial \text{Div}}{\partial z_i^{(k)}} = \frac{\partial \text{Div}}{\partial y_i^{(k)}} \frac{\partial y_i^{(k)}}{\partial z_i^{(k)}} \quad OR \quad \sum_j \frac{\partial \text{Div}}{\partial y_j^{(k)}} \frac{\partial y_j^{(k)}}{\partial z_i^{(k)}} \quad (\text{vector activation})$

- $\frac{\partial \text{Div}}{\partial w_{ji}^{(k+1)}} = y_j^{(k)} \frac{\partial \text{Div}}{\partial z_i^{(k+1)}} \quad \text{for } j = 1 \dots D_{k+1}$

These may be subgradients

# Overall Approach

- For each data instance
  - **Forward pass:** Pass instance forward through the net. Store all intermediate outputs of all computation
  - **Backward pass:** Sweep backward through the net, iteratively compute all derivatives w.r.t weights
- Actual loss is the sum of the divergence over all training instances

$$\textbf{Loss} = \frac{1}{|\{X\}|} \sum_X \text{Div}(Y(X), d(X))$$

- Actual gradient is the sum or average of the derivatives computed for each training instance

$$\nabla_W \textbf{Loss} = \frac{1}{|\{X\}|} \sum_X \nabla_W \text{Div}(Y(X), d(X)) \quad W \leftarrow W - \eta \nabla_W \textbf{Loss}^T$$

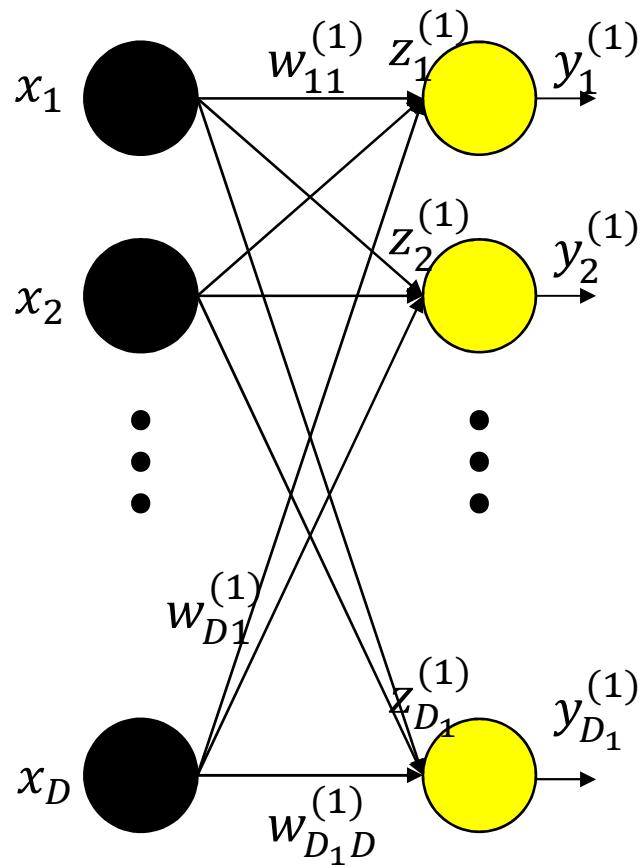
# Training by BackProp

- Initialize weights  $\mathbf{W}^{(k)}$  for all layers  $k = 1 \dots K$
- Do:
  - Initialize  $Loss = 0$ ; For all  $i, j, k$ , initialize  $\frac{dLoss}{dw_{i,j}^{(k)}} = 0$
  - For all  $t = 1:T$  (Loop over training instances)
    - **Forward pass:** Compute
      - Output  $\mathbf{Y}_t$
      - $Loss += Div(\mathbf{Y}_t, \mathbf{d}_t)$
    - **Backward pass:** For all  $i, j, k$ :
      - Compute  $\frac{dDiv(\mathbf{Y}_t, \mathbf{d}_t)}{dw_{i,j}^{(k)}}$
      - Compute  $\frac{dLoss}{dw_{i,j}^{(k)}} += \frac{dDiv(\mathbf{Y}_t, \mathbf{d}_t)}{dw_{i,j}^{(k)}}$
  - For all  $i, j, k$ , update:
$$w_{i,j}^{(k)} = w_{i,j}^{(k)} - \frac{\eta}{T} \frac{dLoss}{dw_{i,j}^{(k)}}$$
- Until  $Loss$  has converged

# Vector formulation

- For layered networks it is generally simpler to think of the process in terms of vector operations
  - Simpler arithmetic
  - Fast matrix libraries make operations *much* faster
- We can restate the entire process in vector terms
  - On slides, please read
  - This is what is *actually* used in any real system
  - Will appear in quiz

# Vector formulation



$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$

$$\mathbf{z}_k = \begin{bmatrix} z_1^{(k)} \\ z_2^{(k)} \\ \vdots \\ z_{D_k}^{(k)} \end{bmatrix}$$

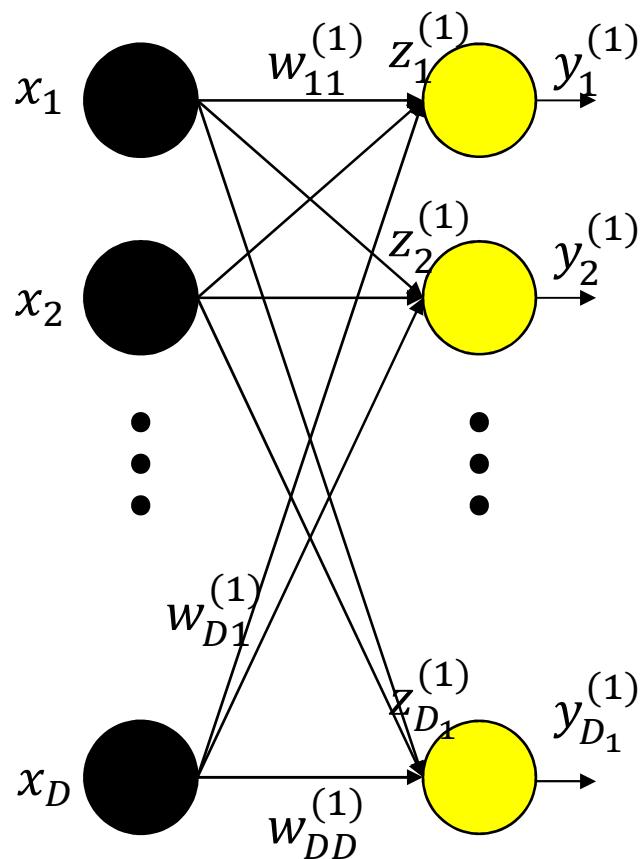
$$\mathbf{y}_k = \begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \vdots \\ y_{D_k}^{(k)} \end{bmatrix}$$

$$\mathbf{W}_k = \begin{bmatrix} w_{11}^{(k)} & w_{21}^{(k)} & \vdots & w_{D_{k-1}1}^{(k)} \\ w_{12}^{(k)} & w_{22}^{(k)} & \vdots & w_{D_{k-1}2}^{(k)} \\ \dots & \dots & \ddots & \vdots \\ w_{1D_k}^{(k)} & w_{2D_k}^{(k)} & \dots & w_{D_{k-1}D_k}^{(k)} \end{bmatrix}$$

$$\mathbf{b}_k = \begin{bmatrix} b_1^{(k)} \\ b_2^{(k)} \\ \vdots \\ b_{D_{k+1}}^{(k)} \end{bmatrix}$$

- Arrange all inputs to the network in a vector  $\mathbf{x}$
- Arrange the *inputs* to neurons of the  $k$ th layer as a vector  $\mathbf{z}_k$
- Arrange the outputs of neurons in the  $k$ th layer as a vector  $\mathbf{y}_k$
- Arrange the weights to any layer as a matrix  $\mathbf{W}_k$ 
  - Similarly with biases

# Vector formulation



$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$

$$\mathbf{z}_k = \begin{bmatrix} z_1^{(k)} \\ z_2^{(k)} \\ \vdots \\ z_{D_k}^{(k)} \end{bmatrix}$$

$$\mathbf{y}_k = \begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \vdots \\ y_{D_k}^{(k)} \end{bmatrix}$$

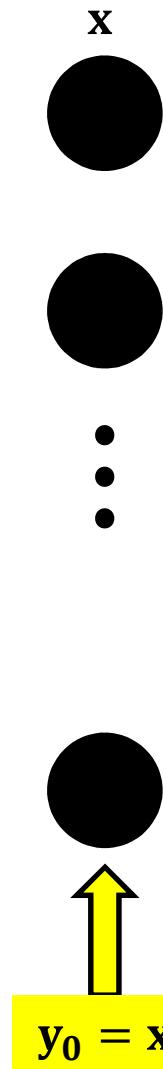
$$\mathbf{W}_k = \begin{bmatrix} w_{11}^{(k)} & w_{21}^{(k)} & \vdots & w_{D_{k-1}1}^{(k)} \\ w_{12}^{(k)} & w_{22}^{(k)} & \vdots & w_{D_{k-1}2}^{(k)} \\ \dots & \dots & \ddots & \vdots \\ w_{1D_k}^{(k)} & w_{2D_k}^{(k)} & \dots & w_{D_{k-1}D_k}^{(k)} \end{bmatrix} \quad \mathbf{b}_k = \begin{bmatrix} b_1^{(k)} \\ b_2^{(k)} \\ \vdots \\ b_{D_{k+1}}^{(k)} \end{bmatrix}$$

- The computation of a single layer is easily expressed in matrix notation as (setting  $\mathbf{y}_0 = \mathbf{x}$ ):

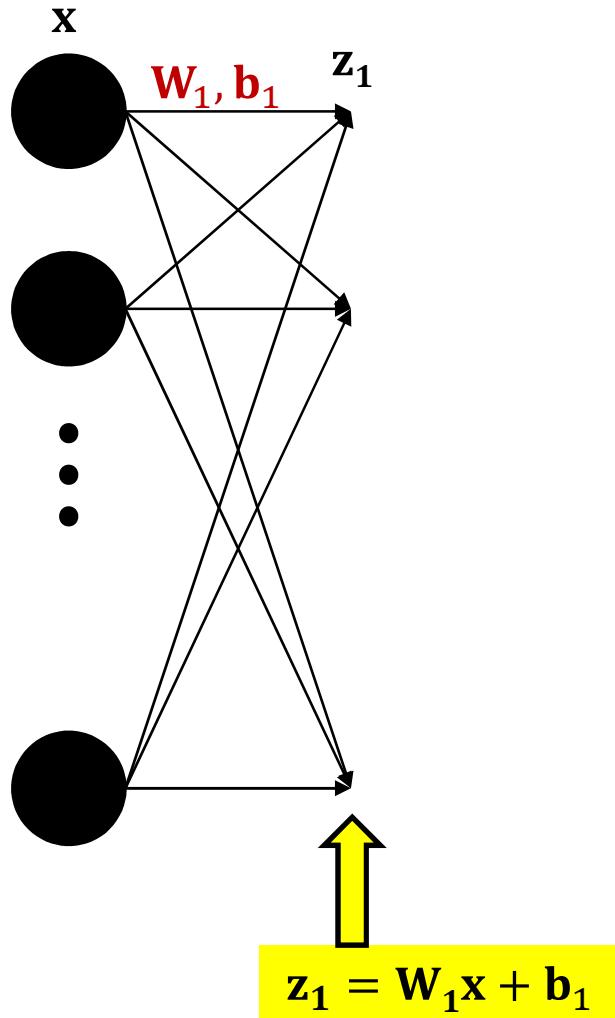
$$\mathbf{z}_k = \mathbf{W}_k \mathbf{y}_{k-1} + \mathbf{b}_k$$

$$\mathbf{y}_k = f_k(\mathbf{z}_k)$$

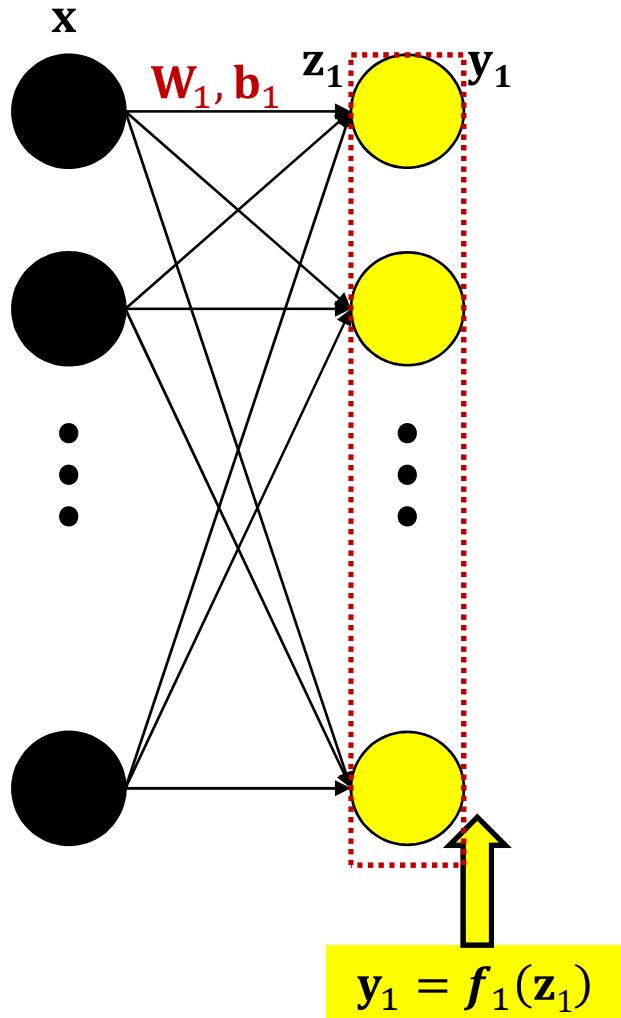
# The forward pass: Evaluating the network



# The forward pass



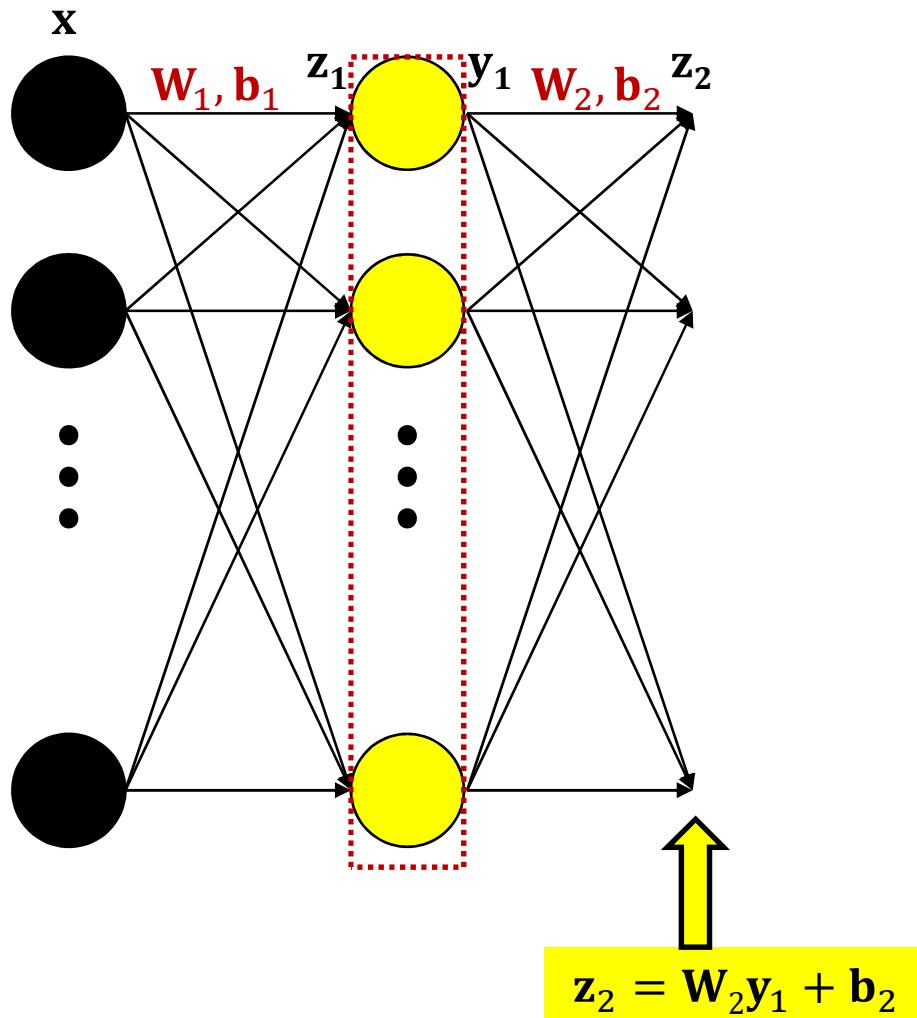
# The forward pass



The Complete computation

$$y_1 = f_1(W_1 x + b_1)$$

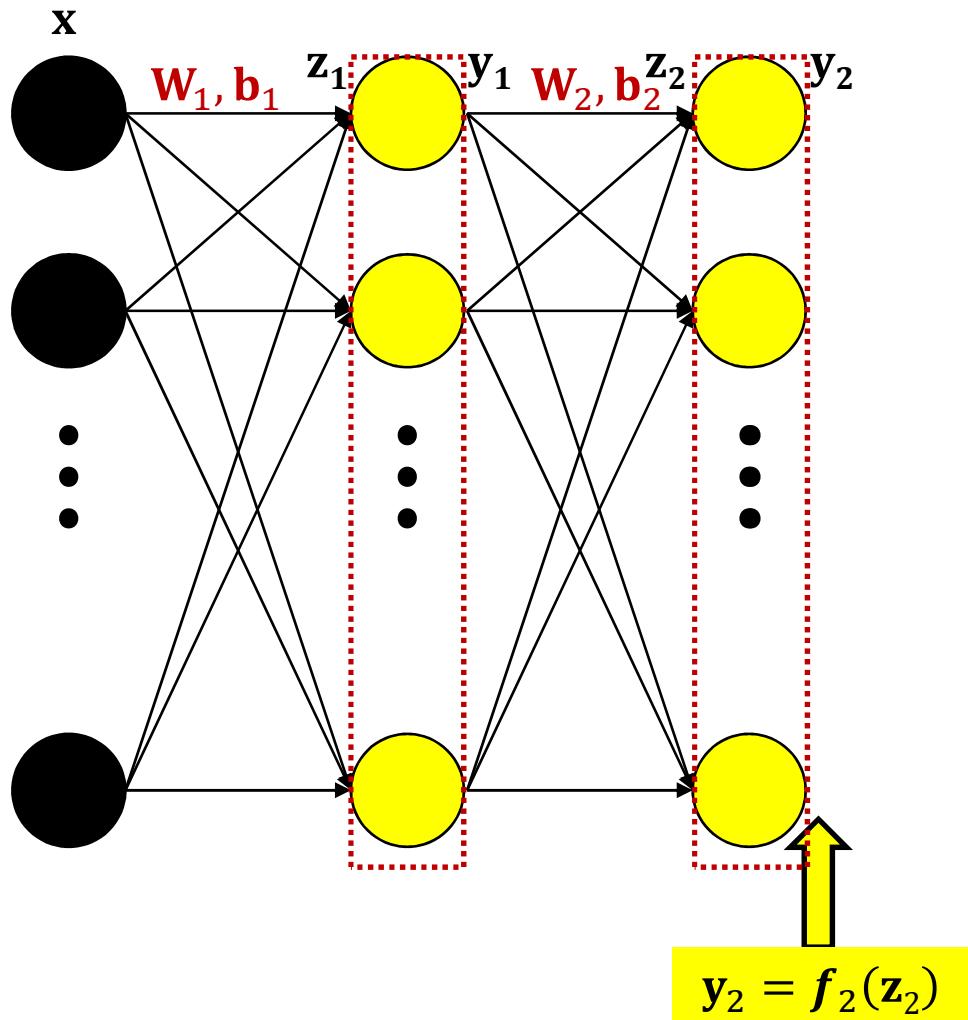
# The forward pass



The Complete computation

$$y_1 = f_1(W_1x + b_1)$$

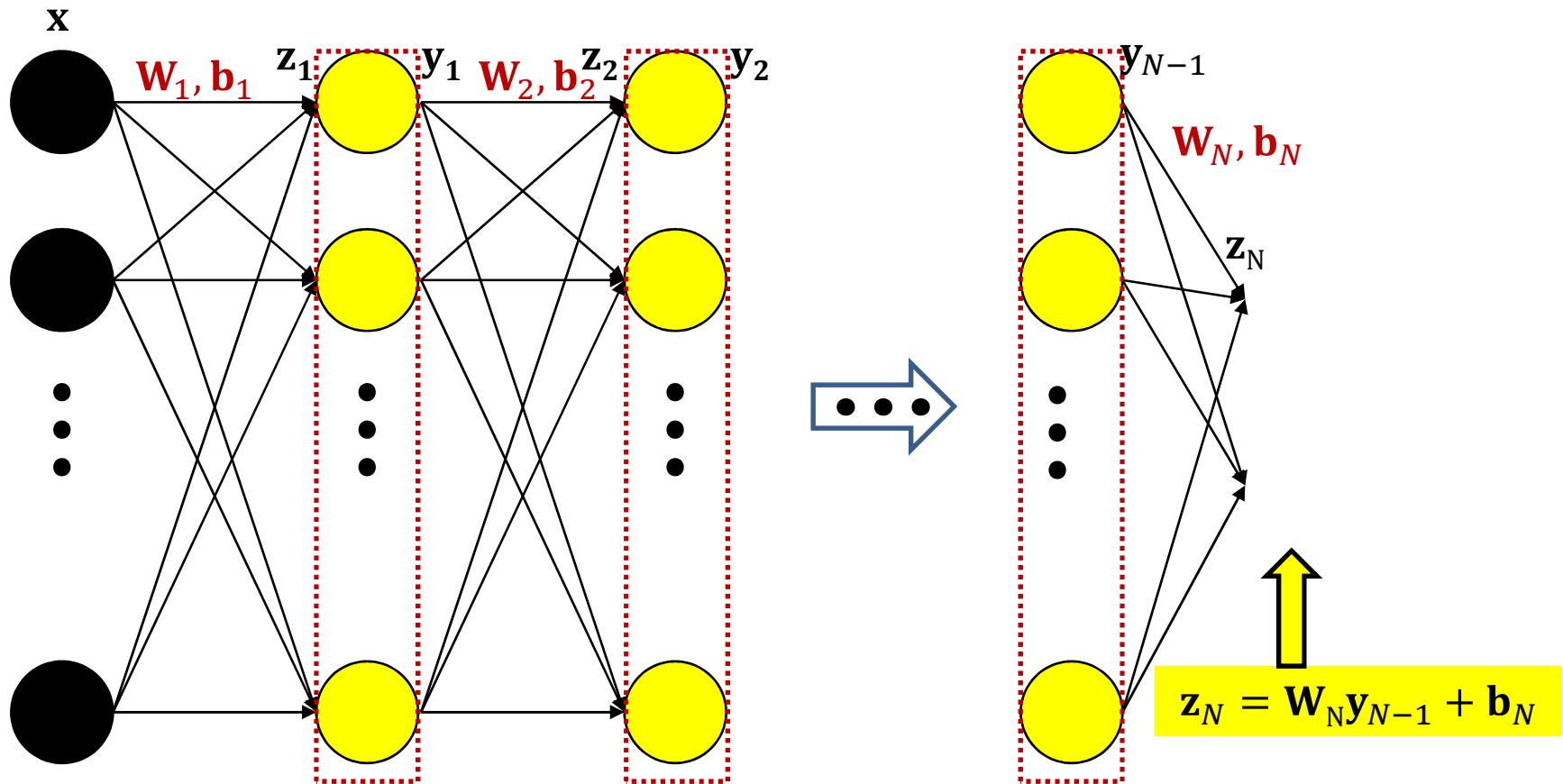
# The forward pass



The Complete computation

$$y_2 = f_2(W_2 f_1(W_1 x + b_1) + b_2)$$

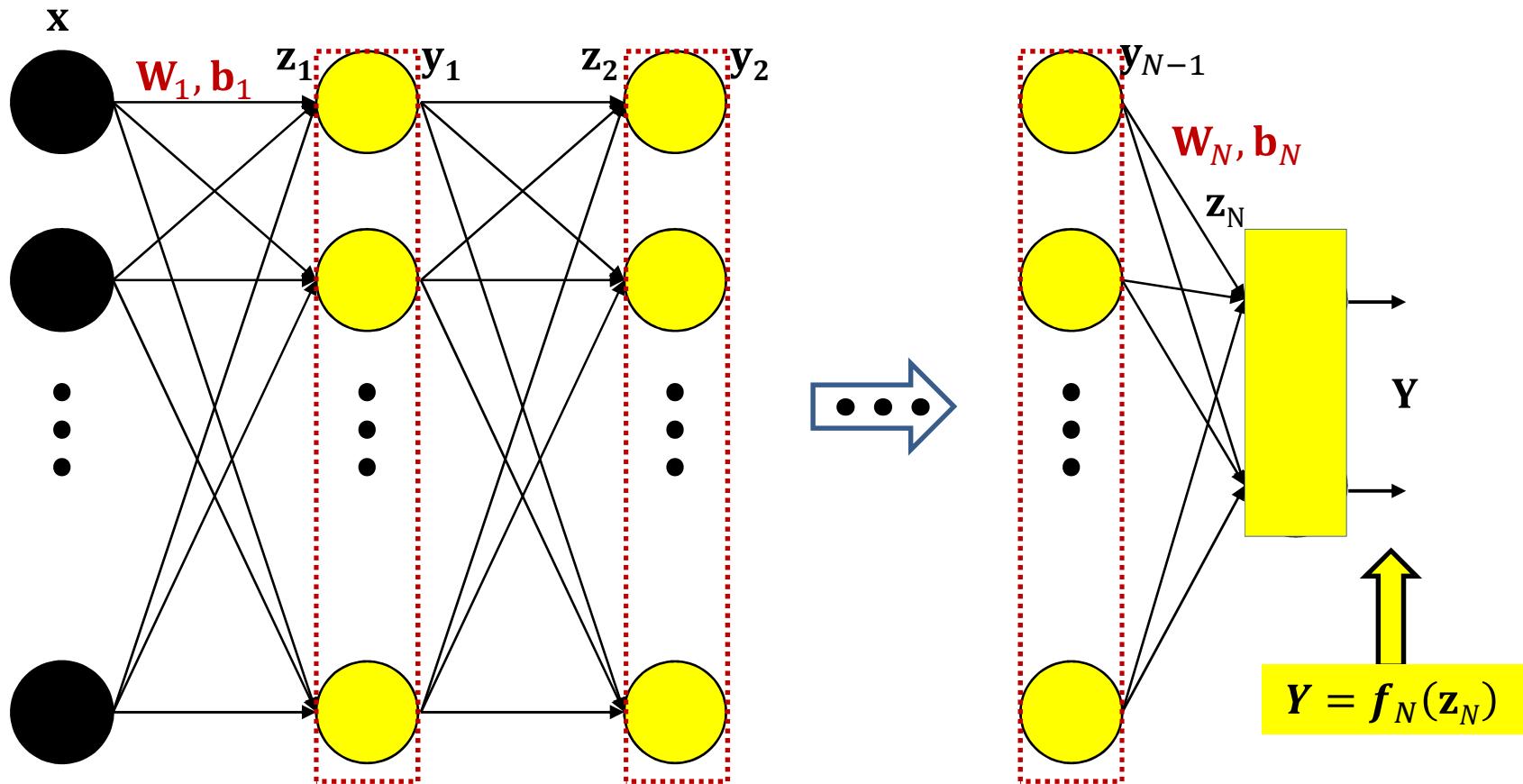
# The forward pass



The Complete computation

$$y_2 = f_2(W_2 f_1(W_1 x + b_1) + b_2)$$

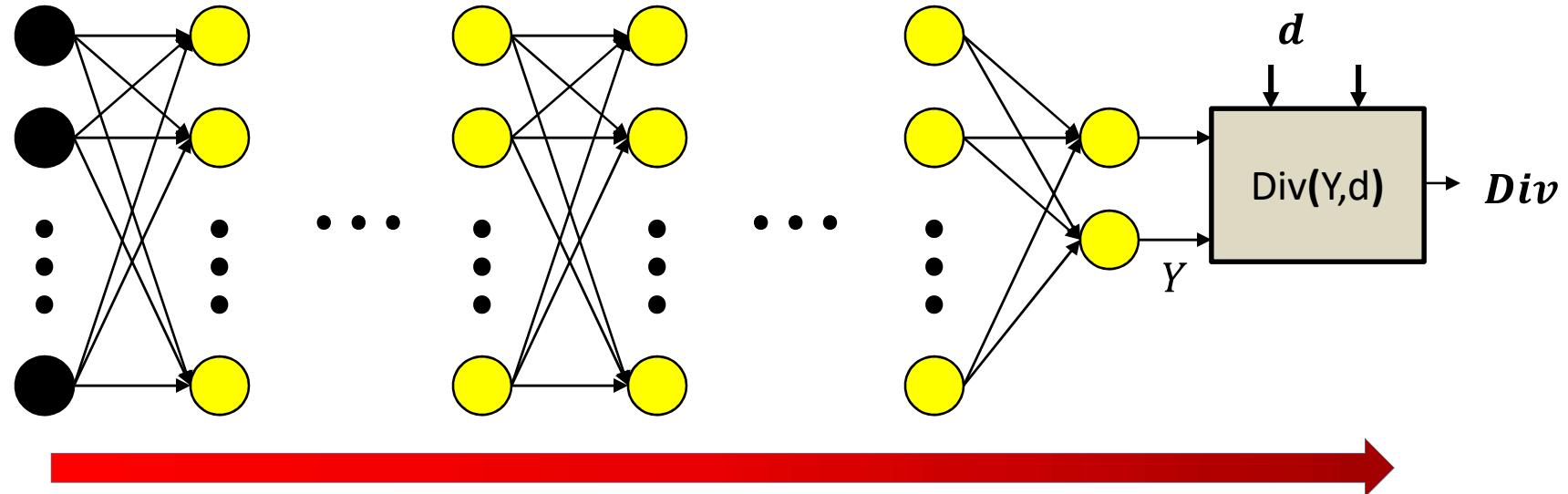
# The forward pass



The Complete computation

$$Y = f_N(\mathbf{W}_N f_{N-1}(\dots f_2(\mathbf{W}_2 f_1(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2) \dots) + \mathbf{b}_N)$$

# Forward pass



## Forward pass:

Initialize

$$\mathbf{y}_0 = \mathbf{x}$$

For  $k = 1$  to  $N$ :

$$\mathbf{z}_k = \mathbf{W}_k \mathbf{y}_{k-1} + \mathbf{b}_k$$

$$\mathbf{y}_k = f_k(\mathbf{z}_k)$$

Output

$$\mathbf{Y} = \mathbf{y}_N$$

# The Forward Pass

- Set  $\mathbf{y}_0 = \mathbf{x}$
- Recursion through layers:
  - For layer  $k = 1$  to  $N$ :

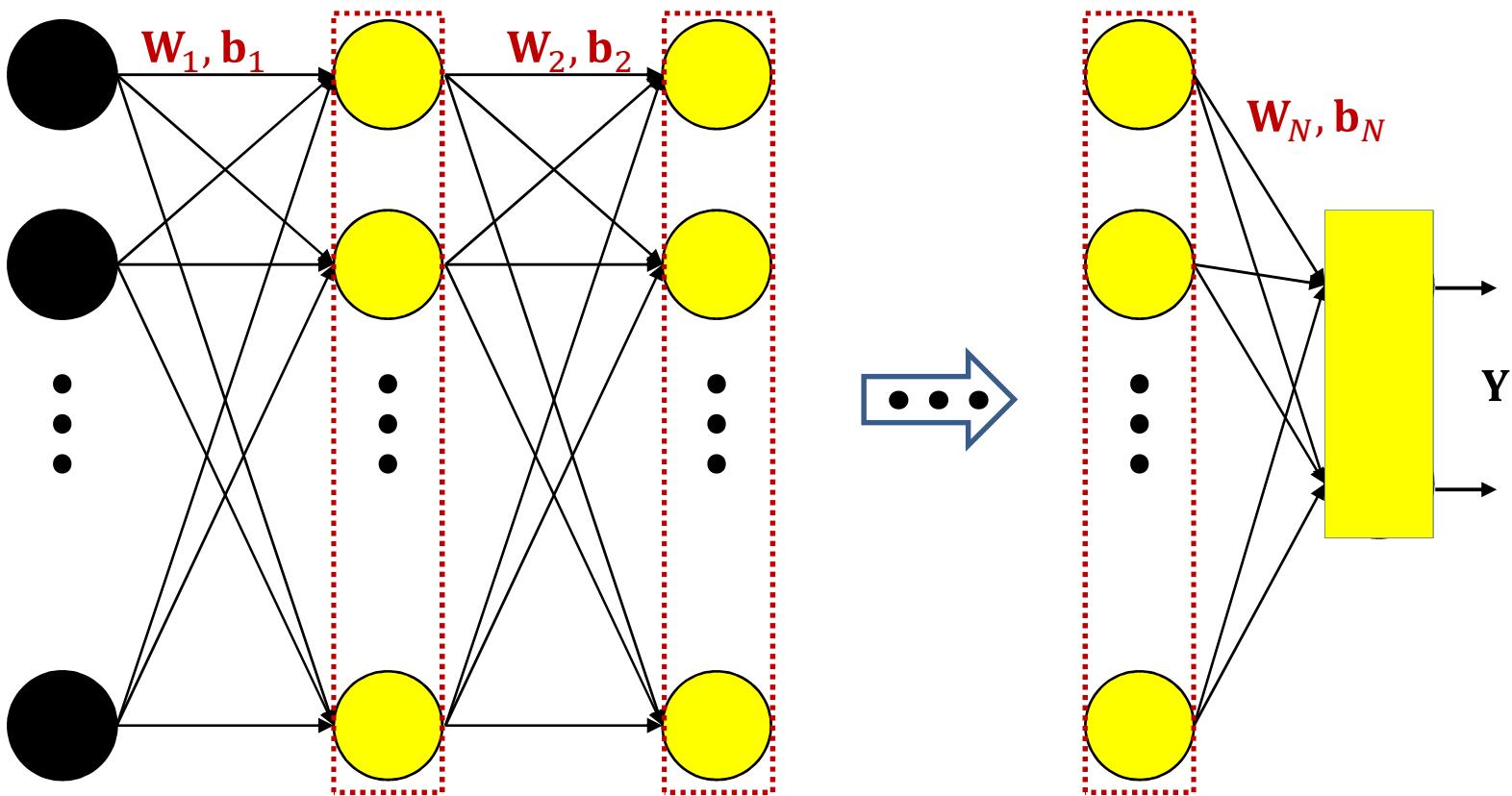
$$\mathbf{z}_k = \mathbf{W}_k \mathbf{y}_{k-1} + \mathbf{b}_k$$

$$\mathbf{y}_k = f_k(\mathbf{z}_k)$$

- Output:

$$\mathbf{Y} = \mathbf{y}_N$$

# The backward pass



- The network is a nested function

$$Y = f_N(\mathbf{W}_N f_{N-1}(\dots f_2(\mathbf{W}_2 f_1(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2) \dots) + \mathbf{b}_N)$$

- The error for any  $\mathbf{x}$  is also a nested function

$$Div(Y, d) = Div(f_N(\mathbf{W}_N f_{N-1}(\dots f_2(\mathbf{W}_2 f_1(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2) \dots) + \mathbf{b}_N), d)$$

# Calculus recap 2: The Jacobian

- The derivative of a vector function w.r.t. vector input is called a *Jacobian*
- It is the matrix of partial derivatives given below

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = f \left( \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_D \end{bmatrix} \right)$$

Using vector notation

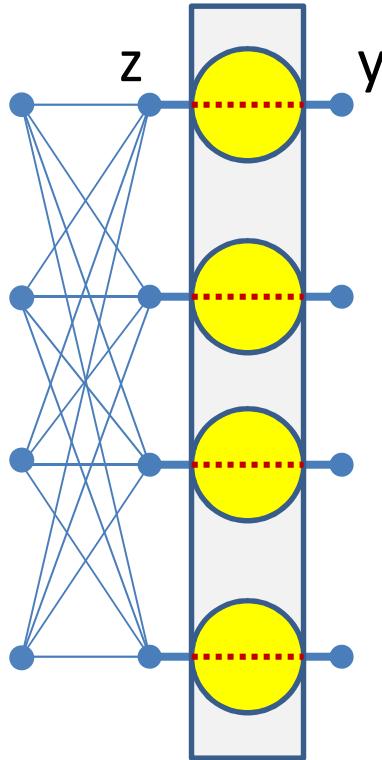
$$\mathbf{y} = f(\mathbf{z})$$

$$J_{\mathbf{y}}(\mathbf{z}) = \begin{bmatrix} \frac{\partial y_1}{\partial z_1} & \frac{\partial y_1}{\partial z_2} & \dots & \frac{\partial y_1}{\partial z_D} \\ \frac{\partial y_2}{\partial z_1} & \frac{\partial y_2}{\partial z_2} & \dots & \frac{\partial y_2}{\partial z_D} \\ \dots & \dots & \ddots & \dots \\ \frac{\partial y_M}{\partial z_1} & \frac{\partial y_M}{\partial z_2} & \dots & \frac{\partial y_M}{\partial z_D} \end{bmatrix}$$

Check:

$$\Delta \mathbf{y} = J_{\mathbf{y}}(\mathbf{z}) \Delta \mathbf{z}$$

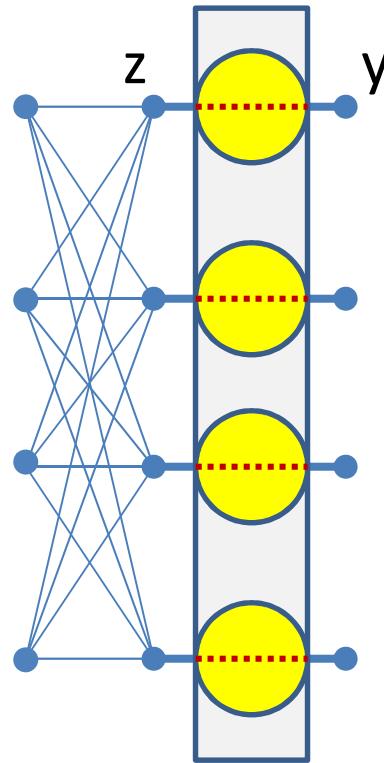
# Jacobians can describe the derivatives of neural activations w.r.t their input



$$J_y(\mathbf{z}) = \begin{bmatrix} \frac{dy_1}{dz_1} & 0 & \dots & 0 \\ 0 & \frac{dy_2}{dz_2} & \dots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & \dots & \frac{dy_D}{dz_D} \end{bmatrix}$$

- **For Scalar activations**
  - Number of outputs is identical to the number of inputs
- Jacobian is a diagonal matrix
  - Diagonal entries are individual derivatives of outputs w.r.t inputs
  - Not showing the superscript “(k)” in equations for brevity

# Jacobians can describe the derivatives of neural activations w.r.t their input

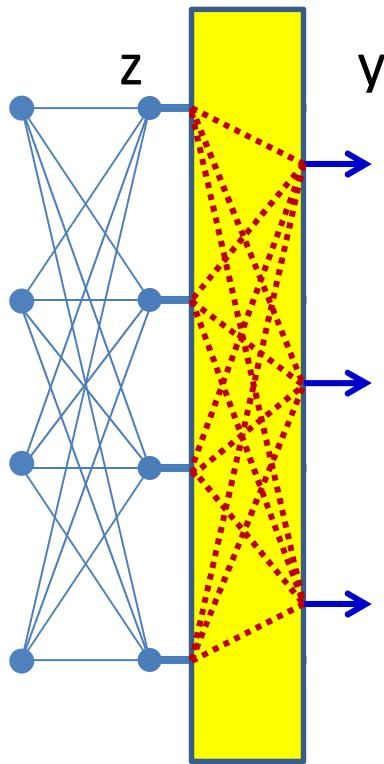


$$y_i = f(z_i)$$

$$J_y(\mathbf{z}) = \begin{bmatrix} f'(y_1) & 0 & \cdots & 0 \\ 0 & f'(y_2) & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & f'(y_M) \end{bmatrix}$$

- **For scalar activations (shorthand notation):**
  - Jacobian is a diagonal matrix
  - Diagonal entries are individual derivatives of outputs w.r.t inputs

# For *Vector* activations



$$J_y(\mathbf{z}) = \begin{bmatrix} \frac{\partial y_1}{\partial z_1} & \frac{\partial y_1}{\partial z_2} & \dots & \frac{\partial y_1}{\partial z_D} \\ \frac{\partial y_2}{\partial z_1} & \frac{\partial y_2}{\partial z_2} & \dots & \frac{\partial y_2}{\partial z_D} \\ \dots & \dots & \ddots & \dots \\ \frac{\partial y_M}{\partial z_1} & \frac{\partial y_M}{\partial z_2} & \dots & \frac{\partial y_M}{\partial z_D} \end{bmatrix}$$

- Jacobian is a full matrix
  - Entries are partial derivatives of individual outputs w.r.t individual inputs

# Special case: Affine functions

$$\mathbf{z} = \mathbf{W}\mathbf{y} + \mathbf{b}$$

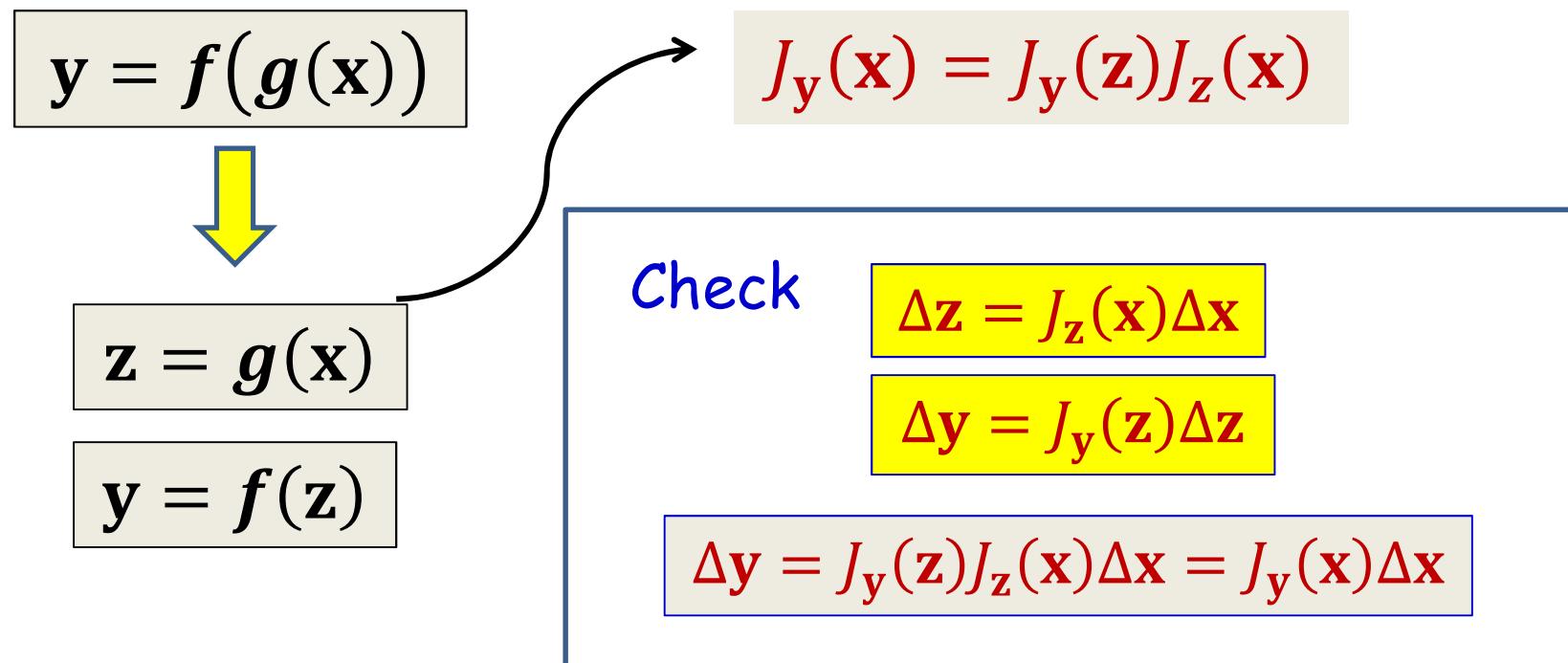


$$J_{\mathbf{z}}(\mathbf{y}) = \mathbf{W}$$

- Matrix  $\mathbf{W}$  and bias  $\mathbf{b}$  operating on vector  $\mathbf{y}$  to produce vector  $\mathbf{z}$
- The Jacobian of  $\mathbf{z}$  w.r.t  $\mathbf{y}$  is simply the matrix  $\mathbf{W}$

# Vector derivatives: Chain rule

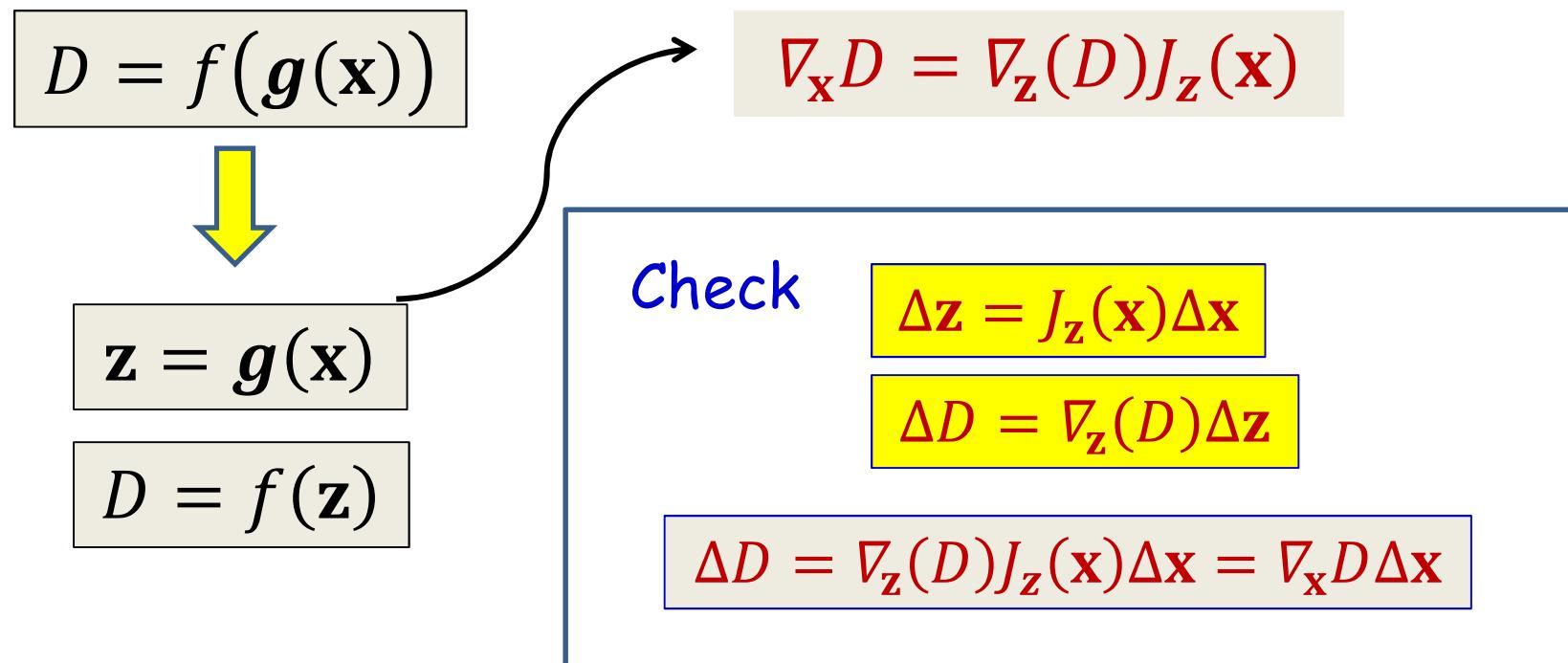
- We can define a chain rule for Jacobians
- **For vector functions of vector inputs:**



Note the order: The derivative of the outer function comes first

# Vector derivatives: Chain rule

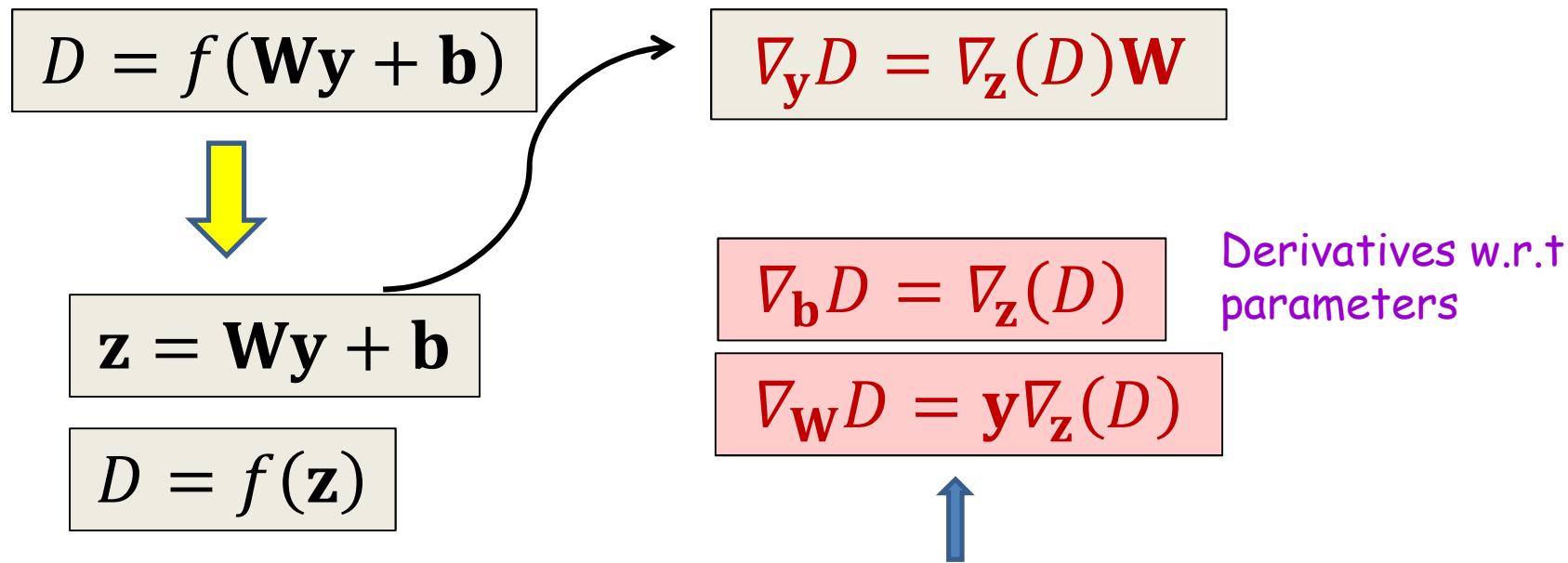
- *The chain rule can combine Jacobians and Gradients*
- **For scalar functions of vector inputs ( $g()$  is vector):**



Note the order: The derivative of the outer function comes first

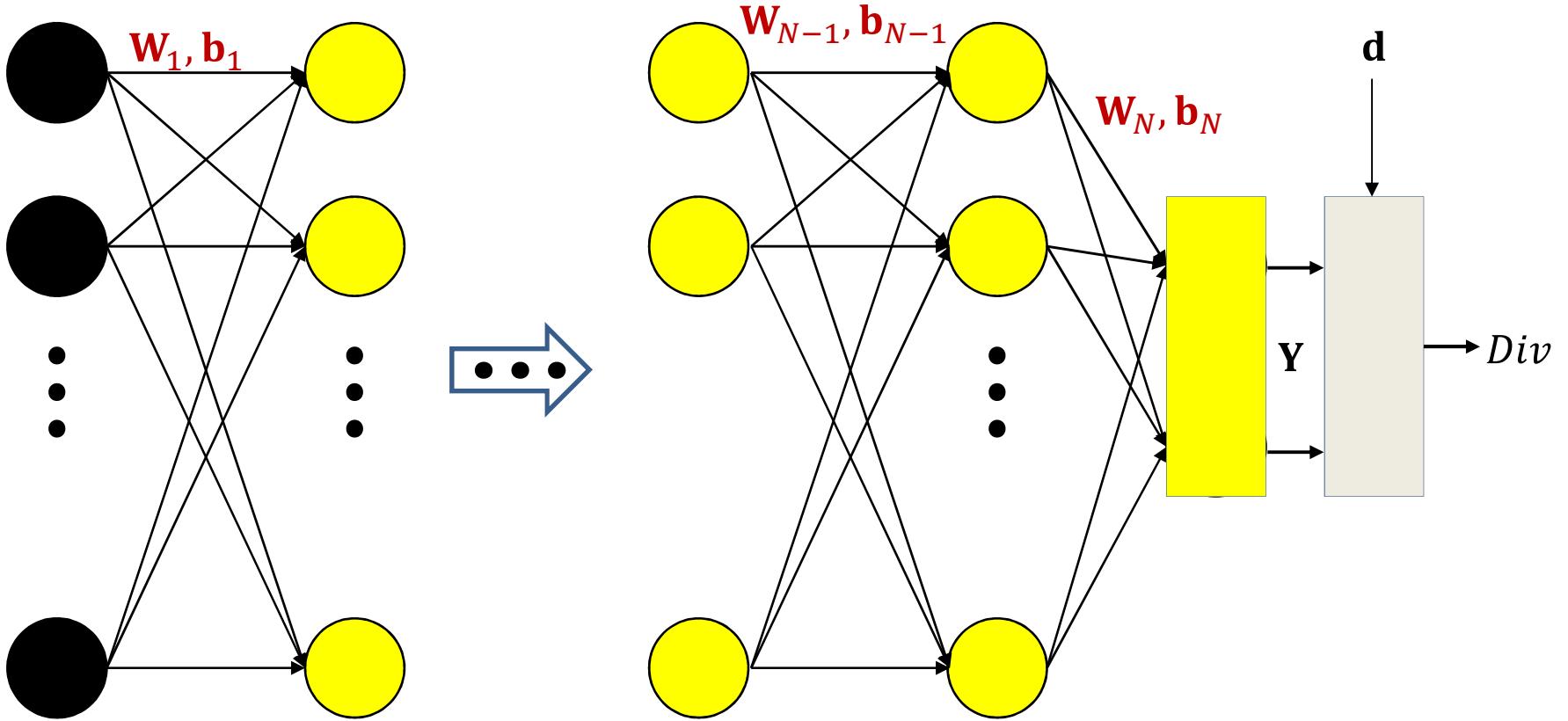
# Special Case

- Scalar functions of Affine functions



Note reversal of order. This is in fact a simplification of a product of tensor terms that occur in the *right* order

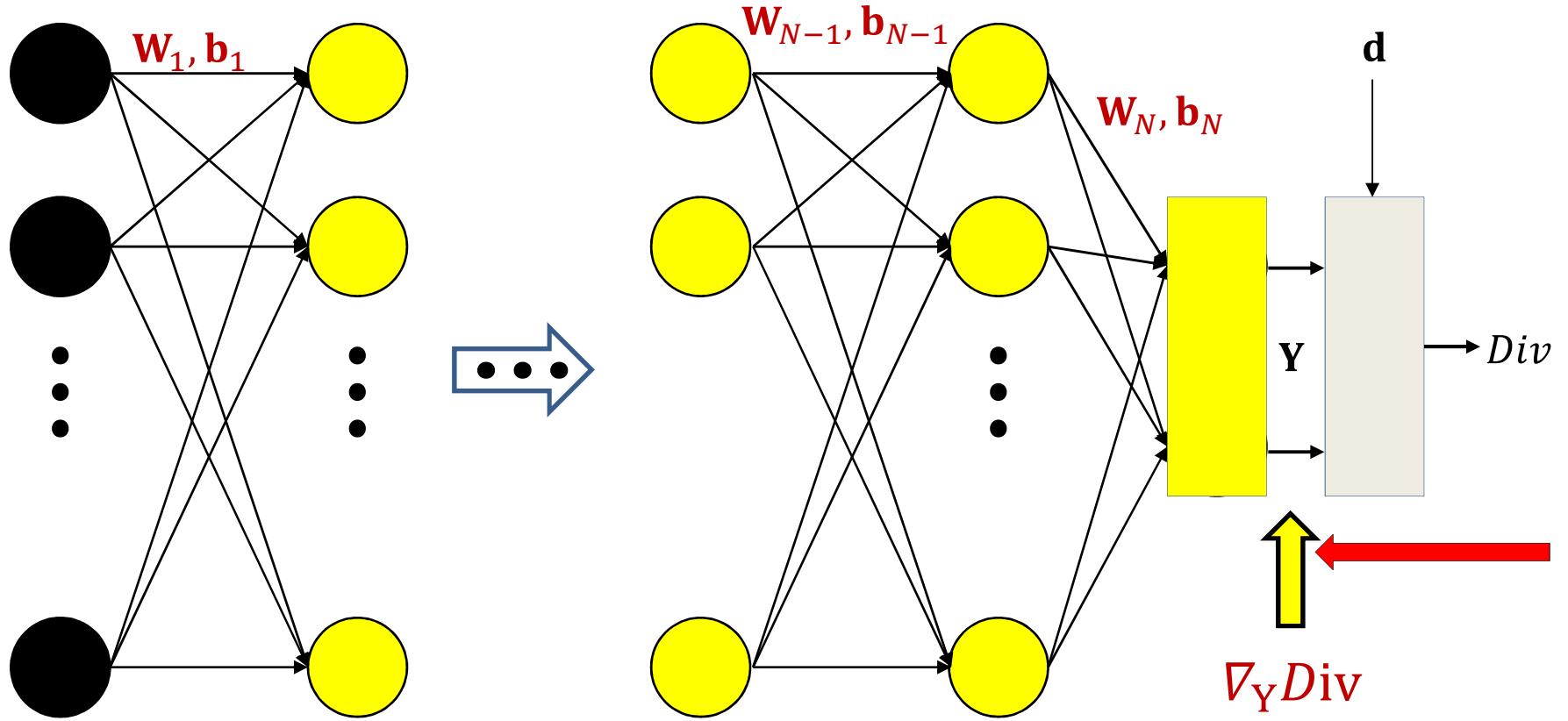
# The backward pass



In the following slides we will also be using the notation  $\nabla_{\mathbf{z}} \mathbf{Y}$  to represent the Jacobian  $J_{\mathbf{Y}}(\mathbf{z})$  to explicitly illustrate the chain rule

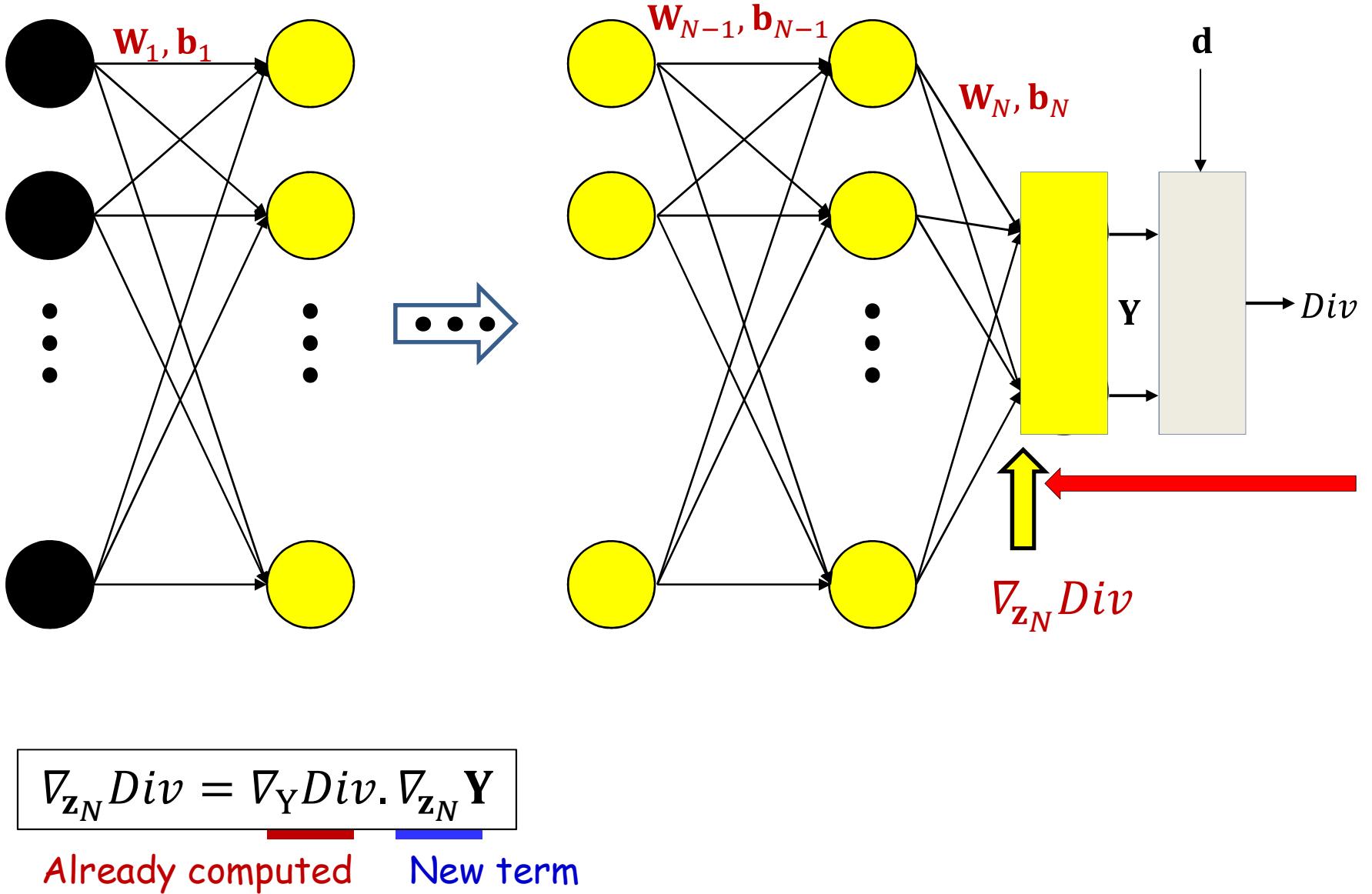
In general  $\nabla_{\mathbf{a}} \mathbf{b}$  represents a derivative of  $\mathbf{b}$  w.r.t.  $\mathbf{a}$  and could be a the transposed gradient (for scalar  $\mathbf{b}$ ) or a Jacobian (for vector  $\mathbf{b}$ )

# The backward pass

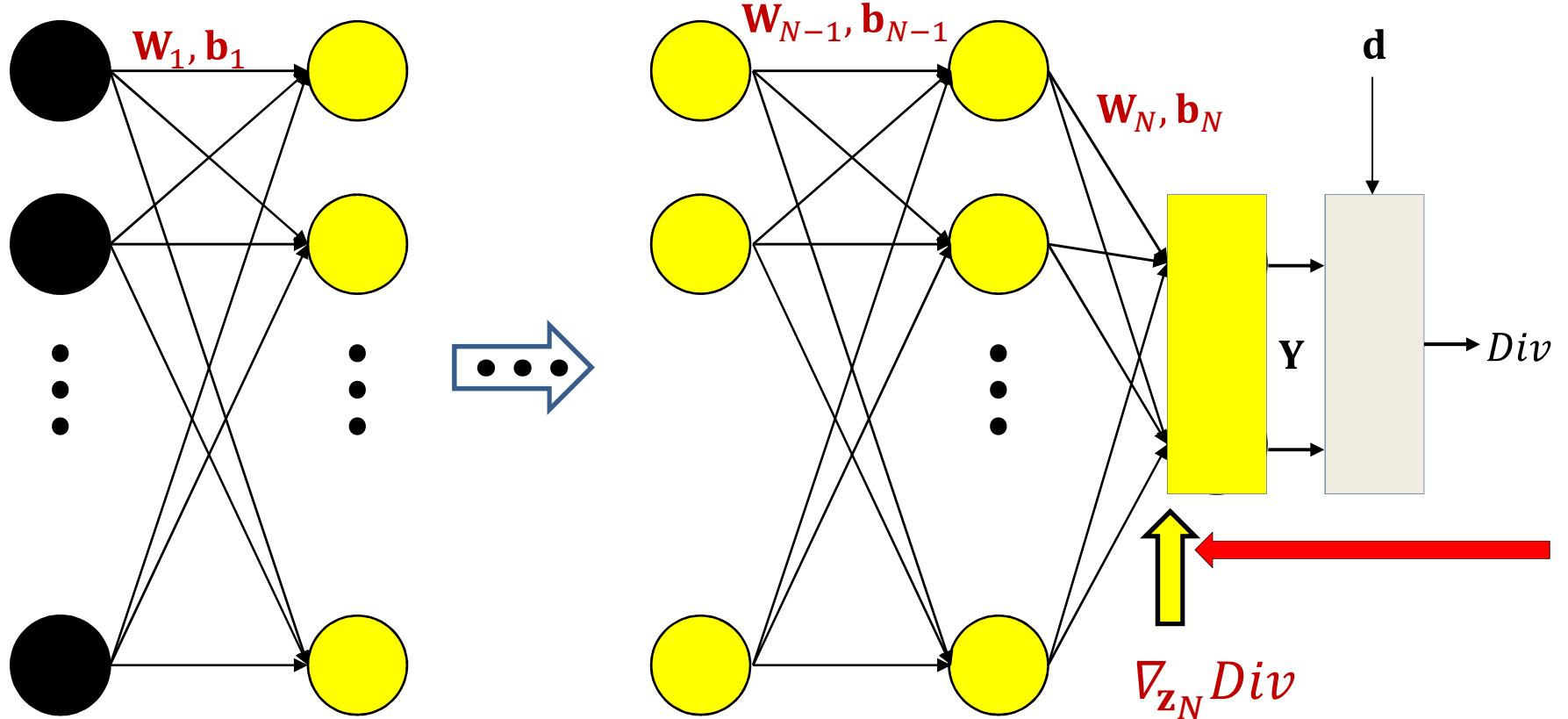


First compute the gradient of the divergence w.r.t.  $Y$ .  
The actual gradient depends on the divergence function.

# The backward pass



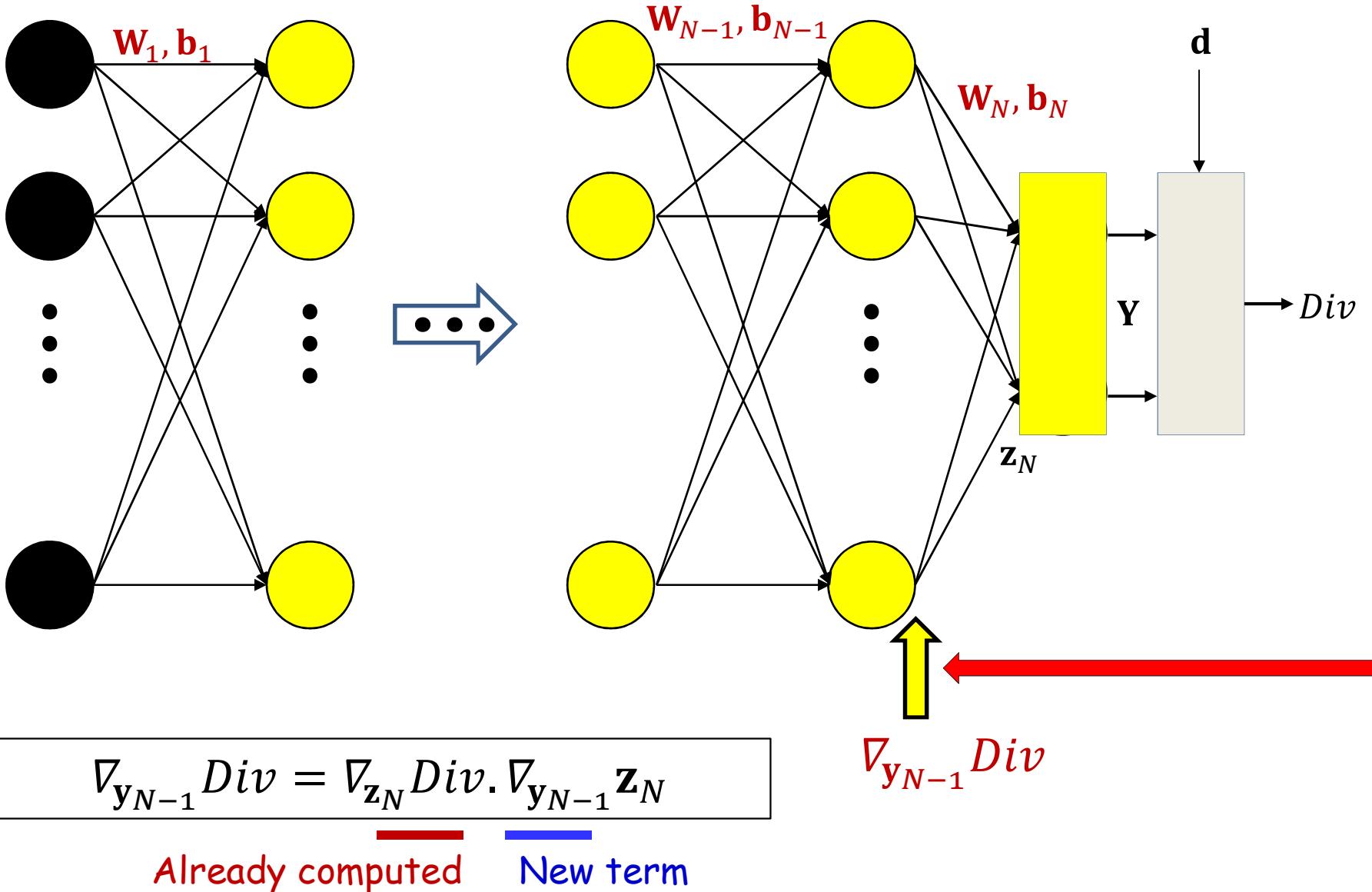
# The backward pass



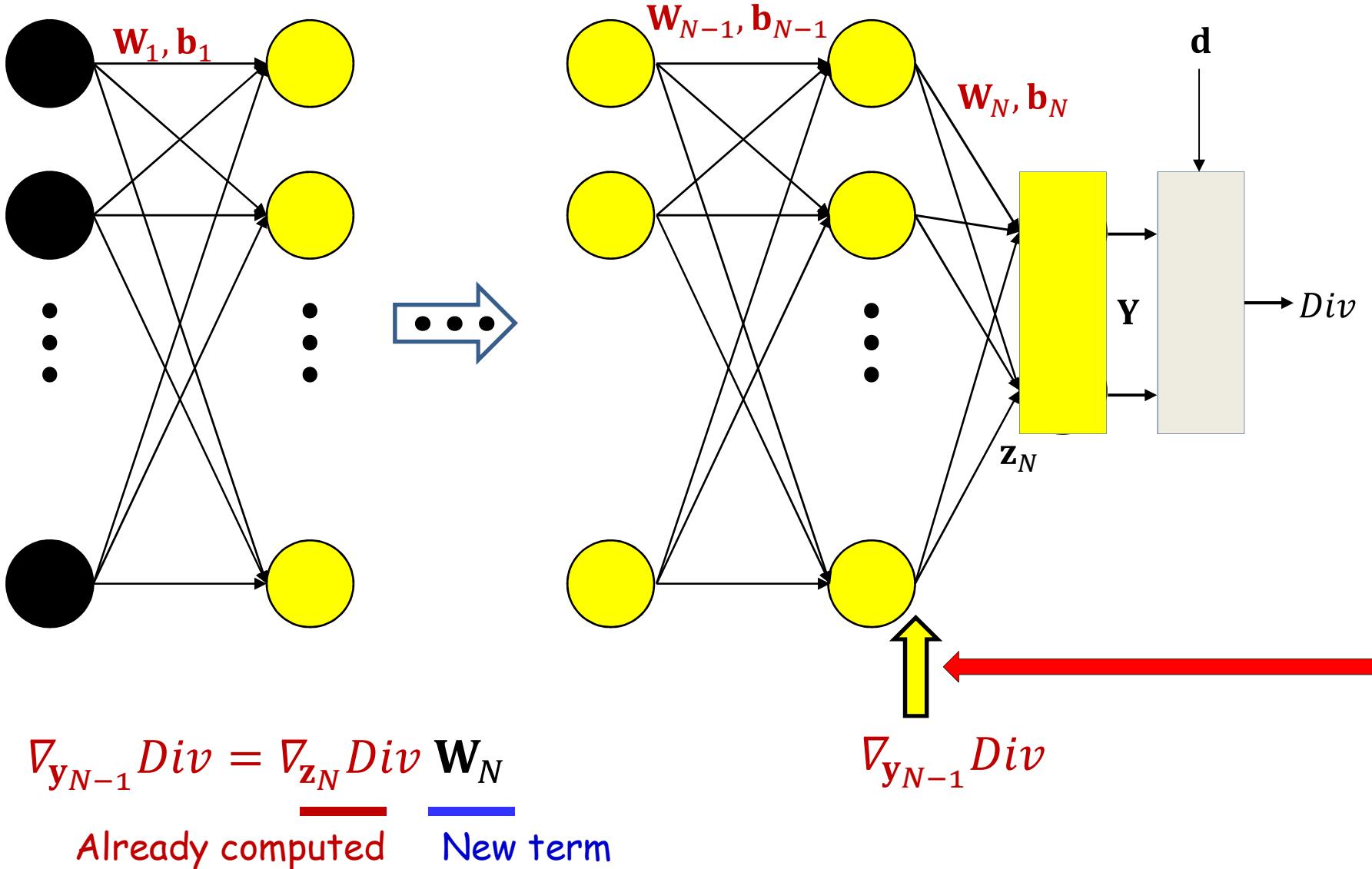
$$\nabla_{\mathbf{z}_N} Div = \underline{\nabla_{\mathbf{Y}} Div} \underline{J_{\mathbf{Y}}(\mathbf{z}_N)}$$

Already computed      New term

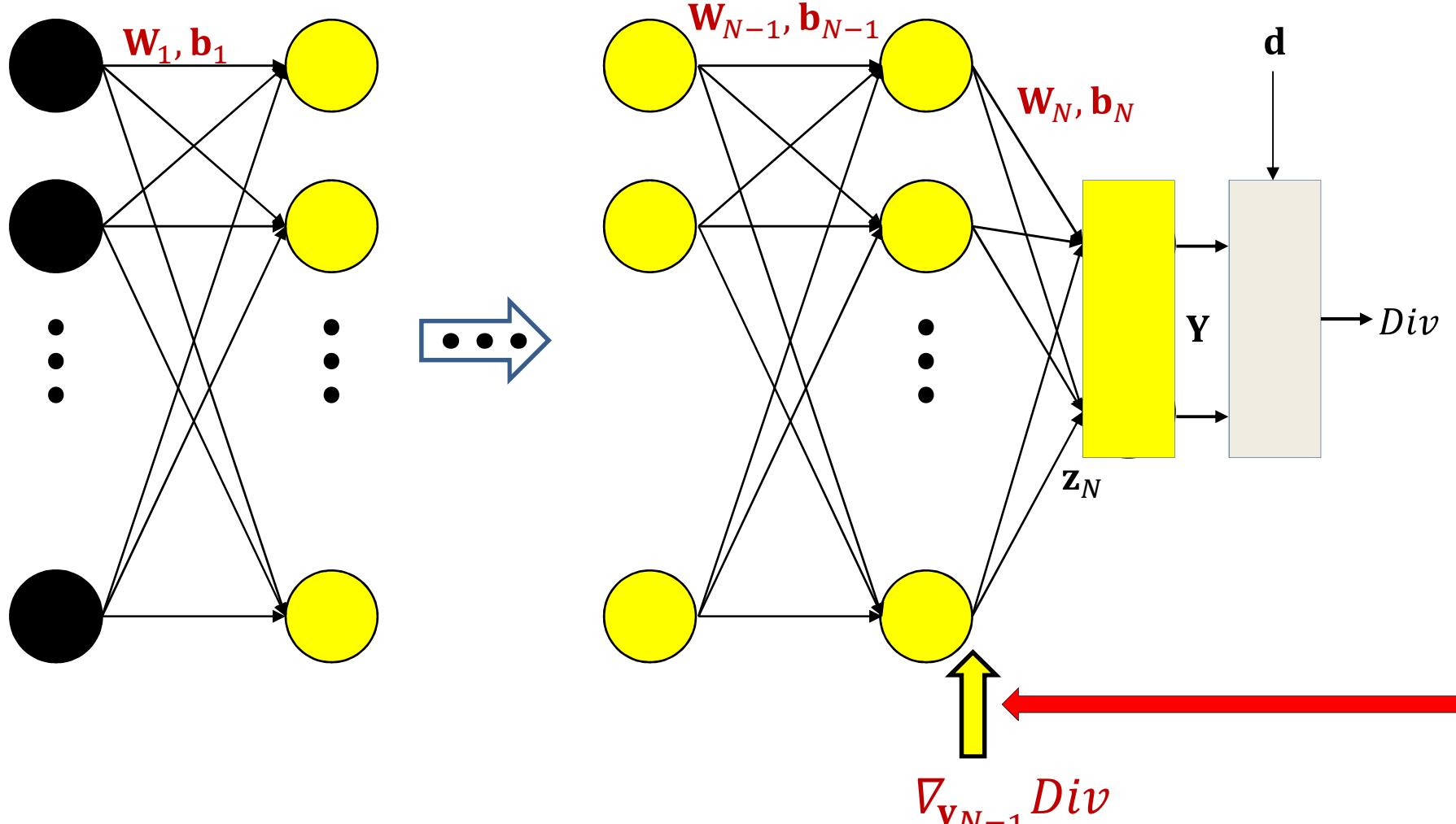
# The backward pass



# The backward pass



# The backward pass

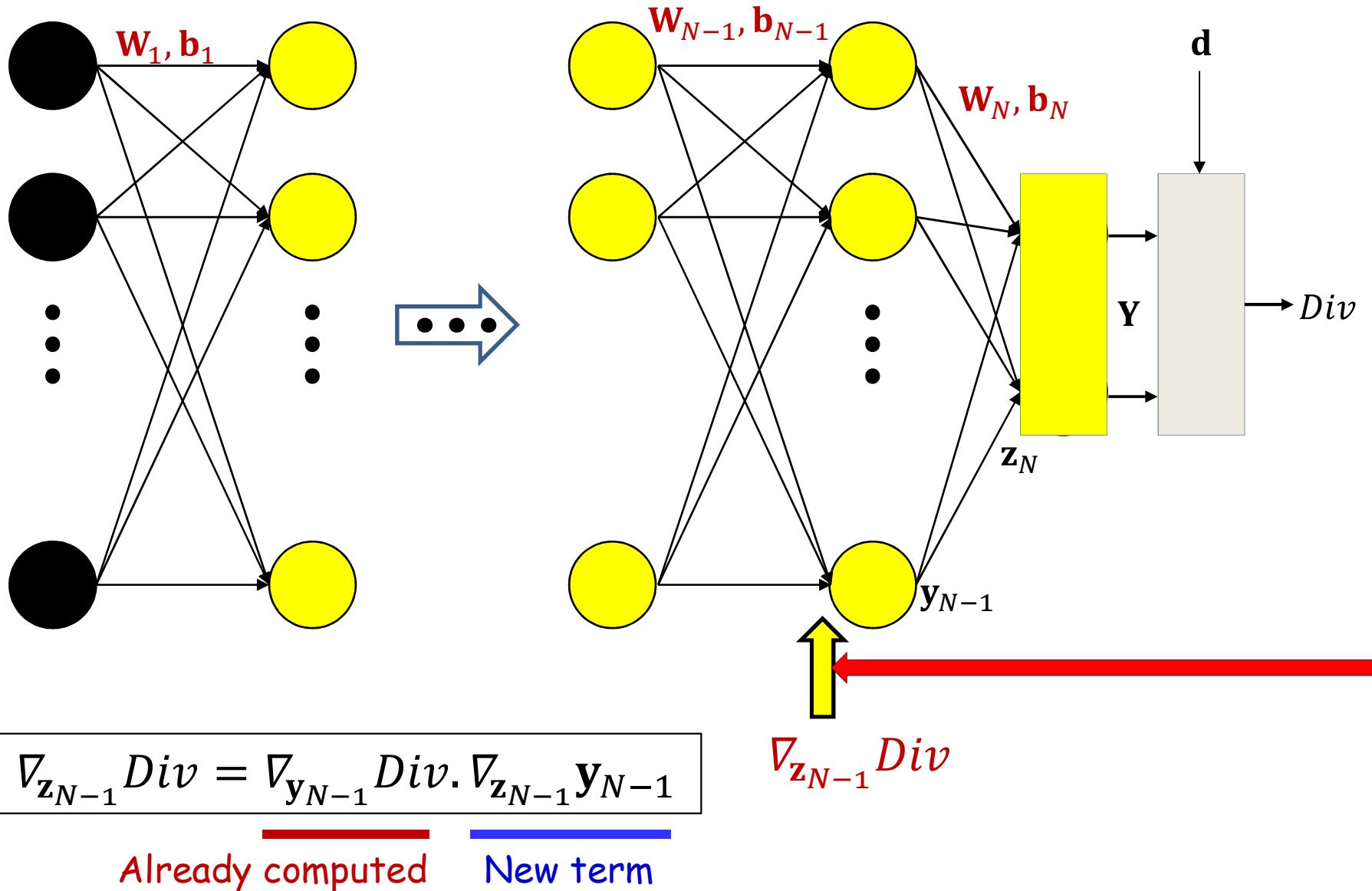


$$\nabla_{y_{N-1}} Div = \nabla_{z_N} Div \mathbf{W}_N$$

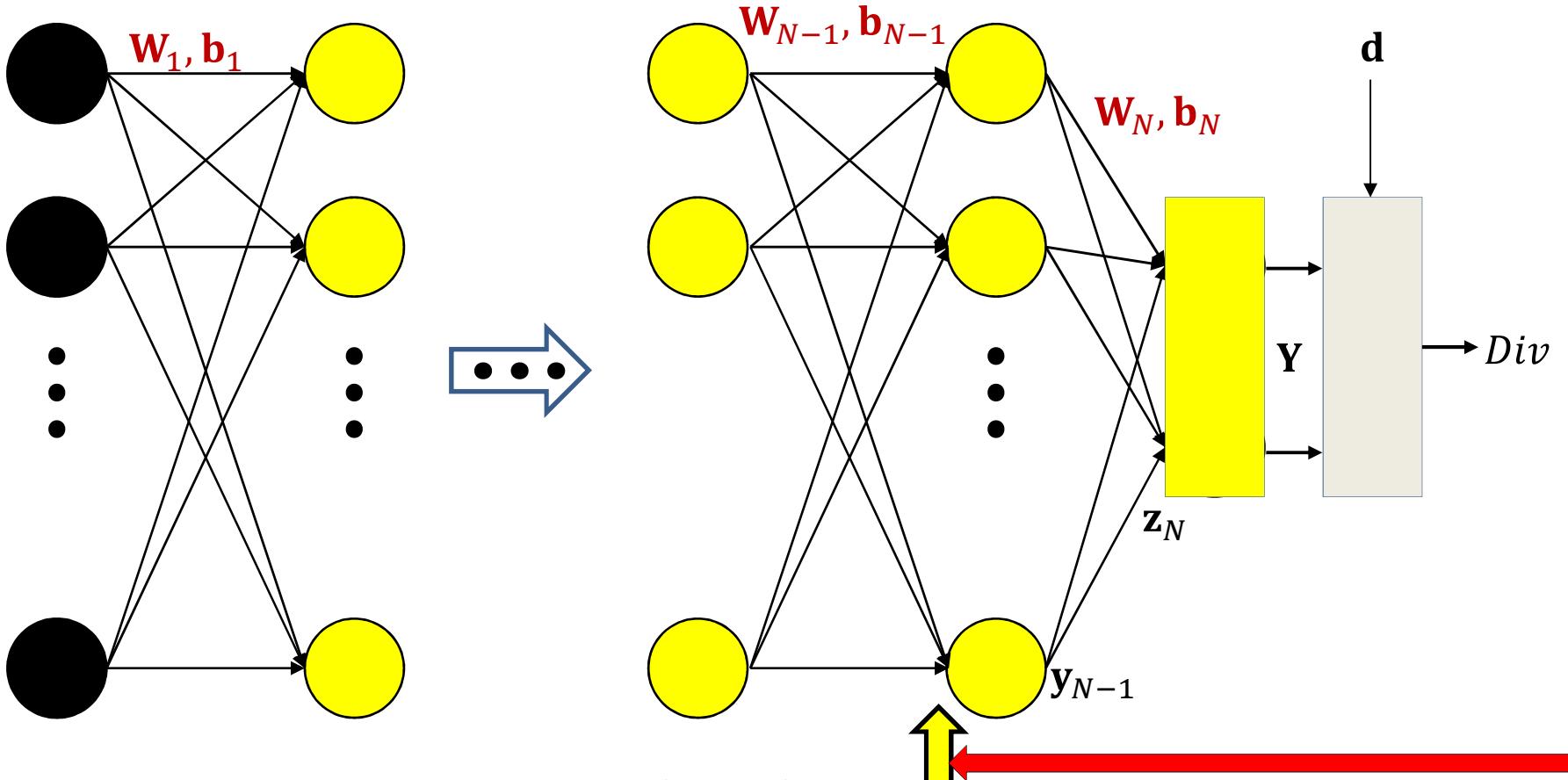
$$\nabla_{\mathbf{W}_N} Div = \mathbf{y}_{N-1} \nabla_{z_N} Div$$

$$\nabla_{\mathbf{b}_N} Div = \nabla_{z_N} Div$$

# The backward pass



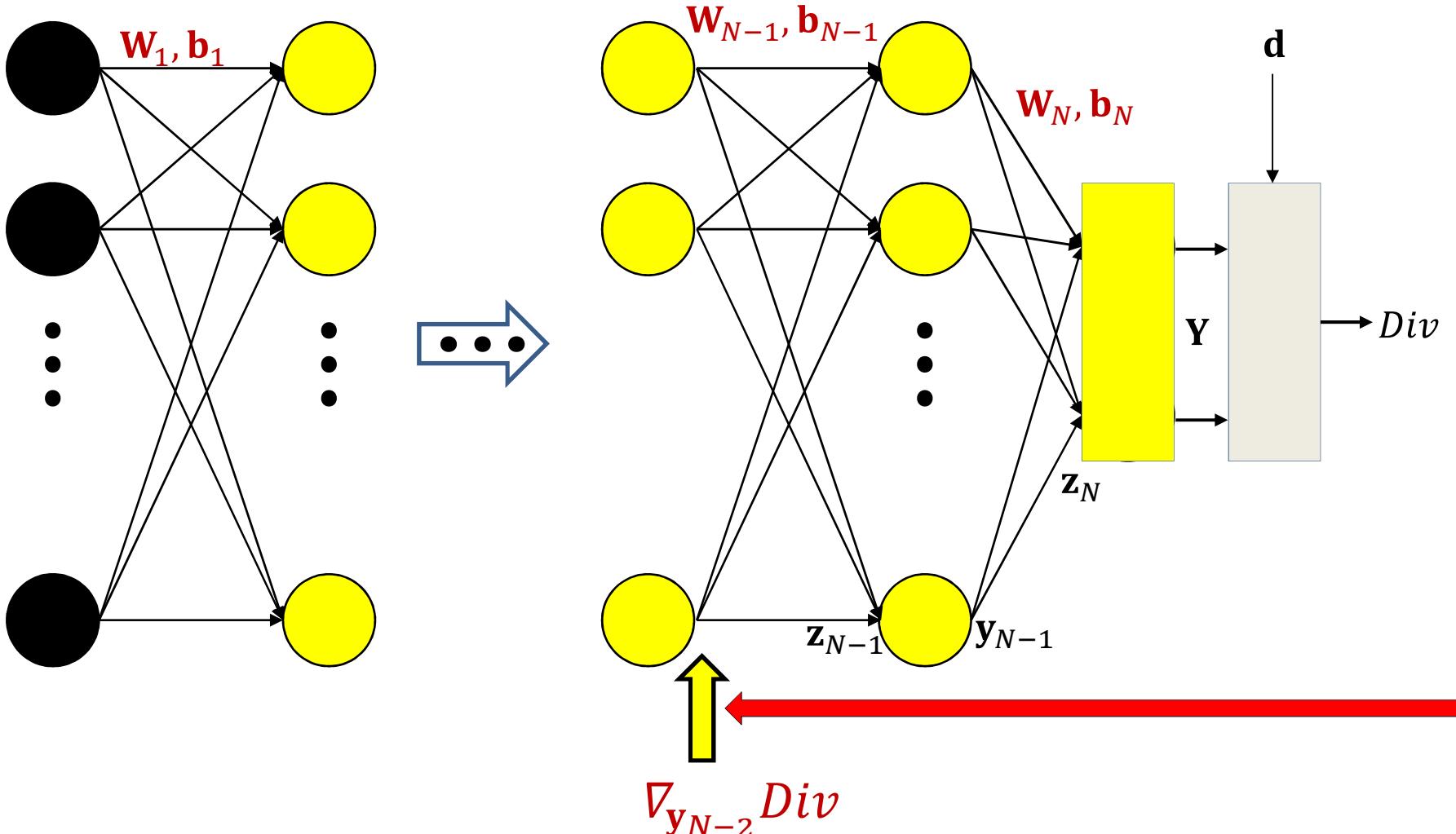
# The backward pass



$$\nabla_{z_{N-1}} \text{Div} = \nabla_{y_{N-1}} \text{Div} J_{y_{N-1}}(z_{N-1})$$

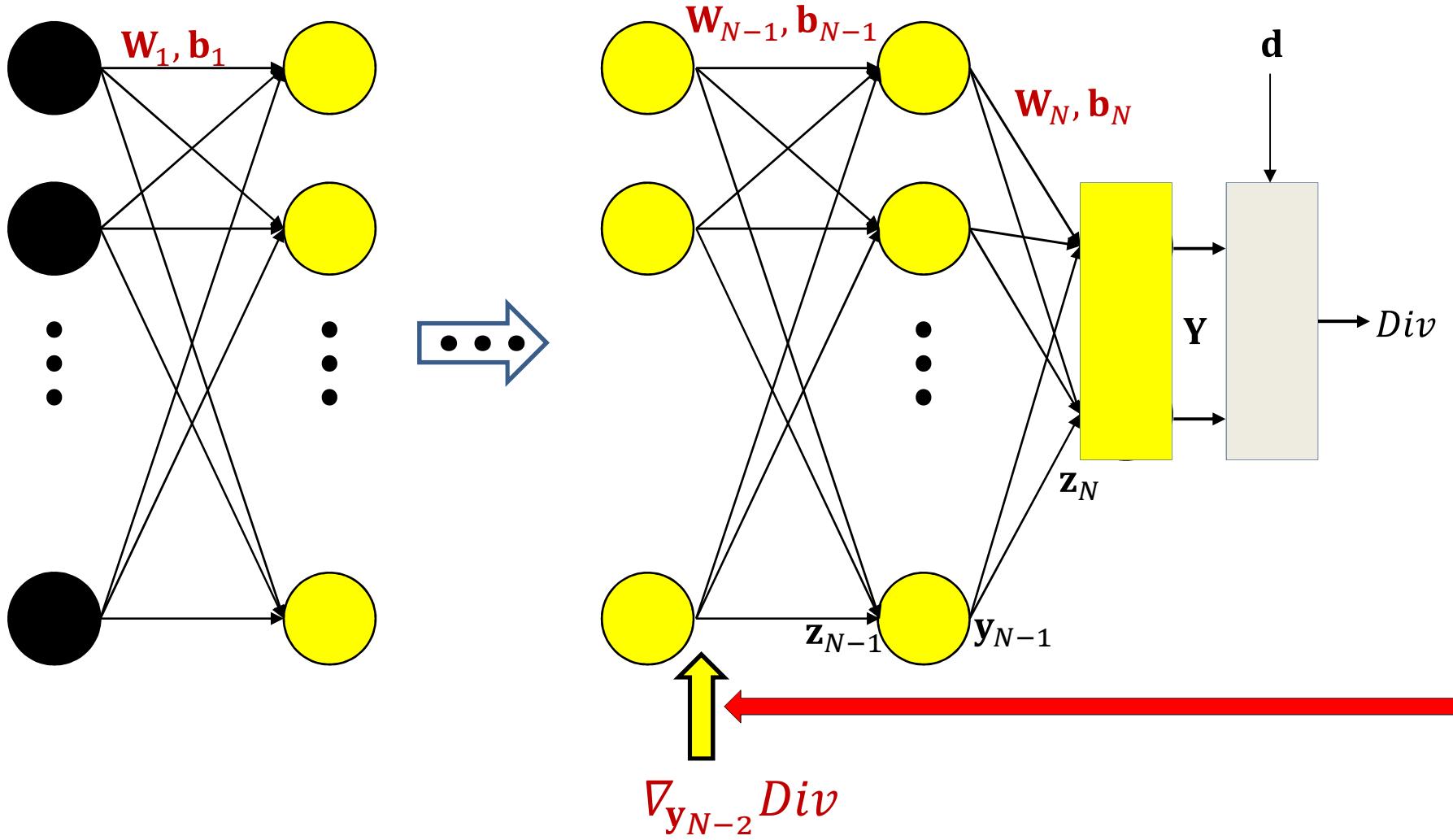
The Jacobian will be a diagonal matrix for scalar activations

# The backward pass



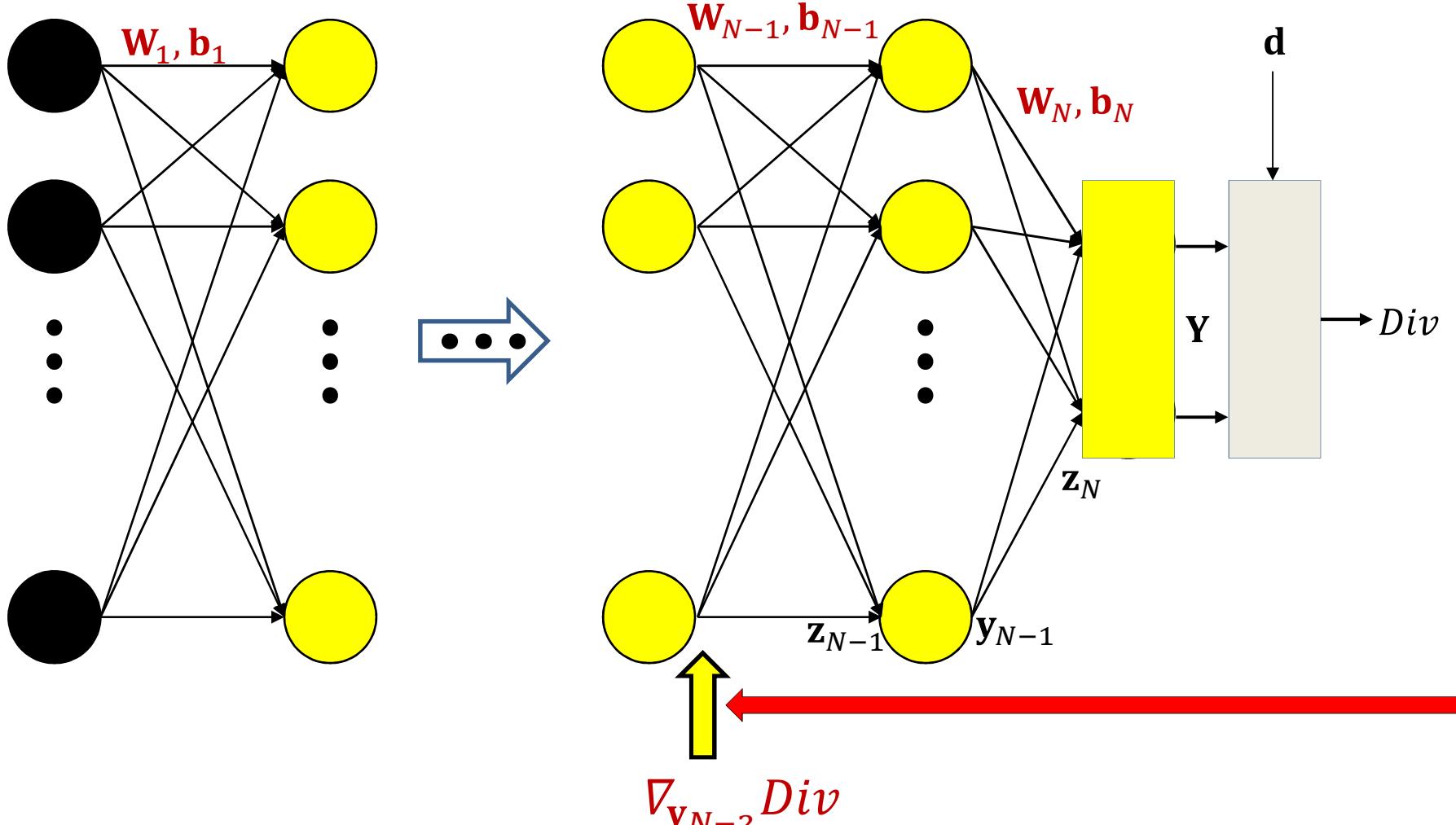
$$\nabla_{\mathbf{y}_{N-2}} Div = \nabla_{\mathbf{z}_{N-1}} Div \cdot \nabla_{\mathbf{y}_{N-2}} \mathbf{z}_{N-1}$$

# The backward pass



$$\nabla_{y_{N-2}} Div = \nabla_{z_{N-1}} Div \mathbf{W}_{N-1}$$

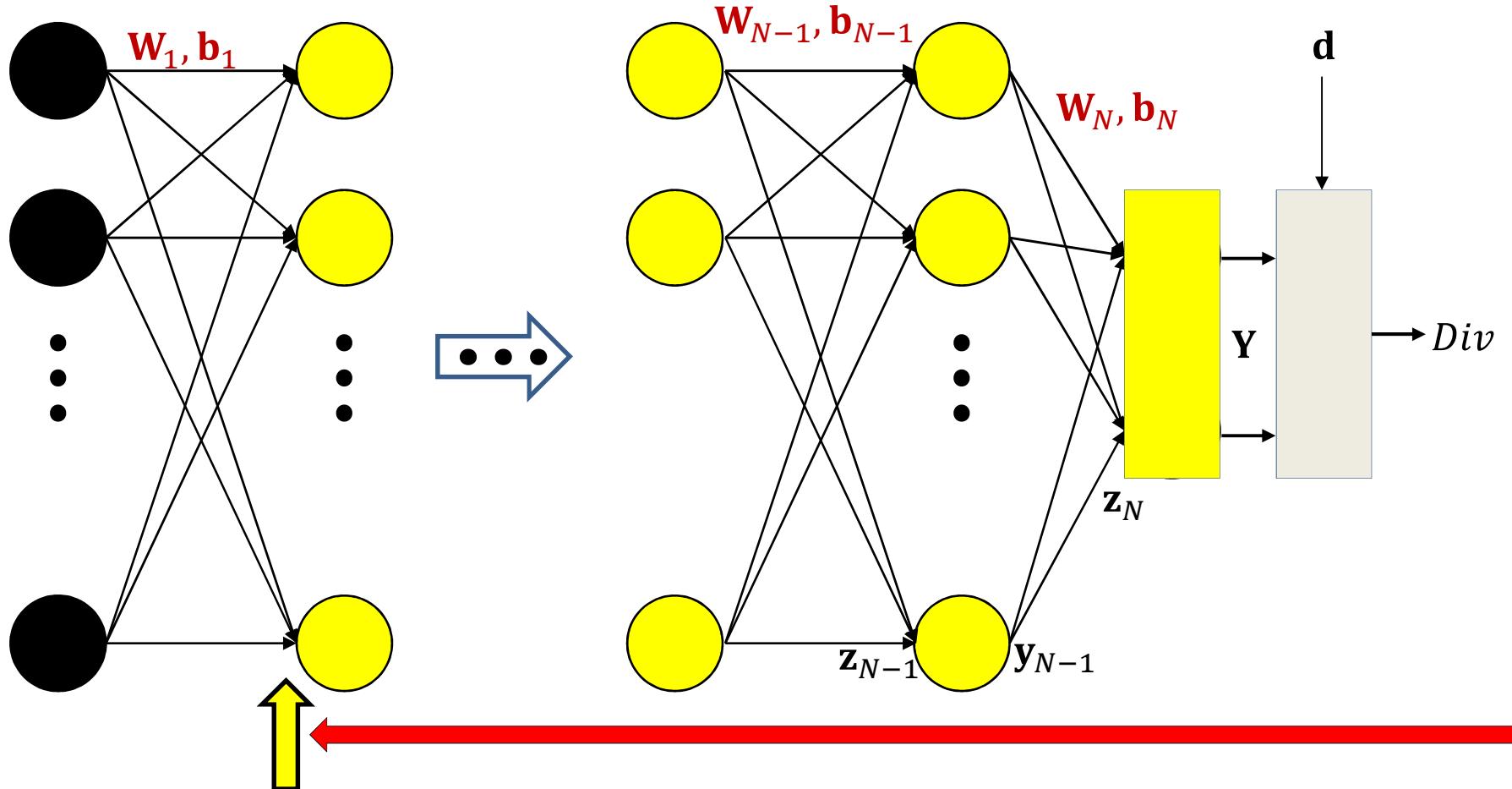
# The backward pass



$$\nabla_{y_{N-2}} Div = \nabla_{z_{N-1}} Div \quad W_{N-1}$$

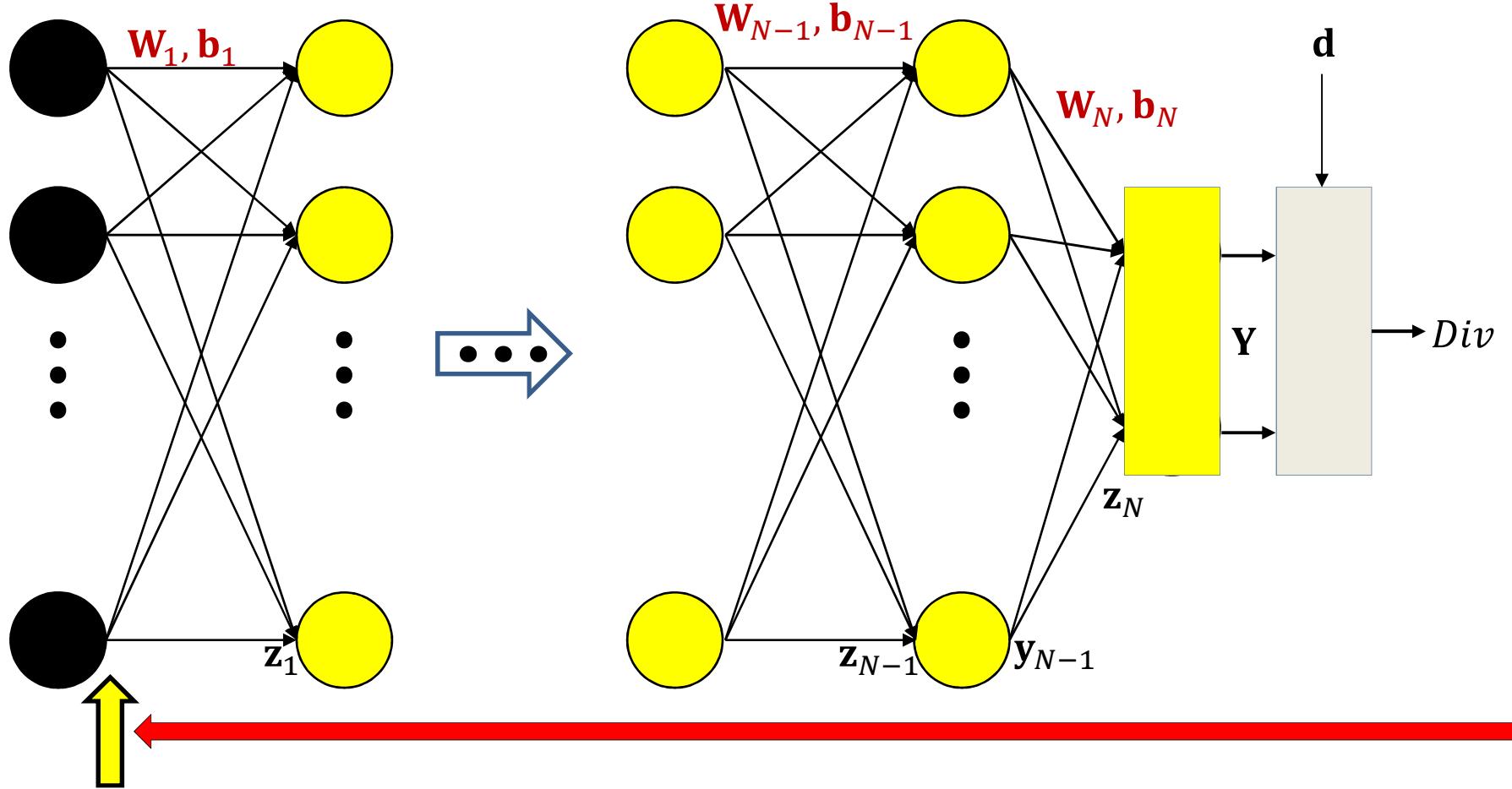
|   |
|---|
| $\nabla_{W_{N-1}} Div = y_{N-2} \nabla_{z_{N-1}} Div$ |
| $\nabla_{b_{N-1}} Div = \nabla_{z_{N-1}} Div$         |

# The backward pass



$$\nabla_{\mathbf{z}_1} \text{Div} = \nabla_{\mathbf{y}_1} \text{Div} J_{\mathbf{y}_1}(\mathbf{z}_1)$$

# The backward pass



$$\nabla_{\mathbf{W}_1} \text{Div} = \mathbf{x} \nabla_{\mathbf{z}_1} \text{Div}$$

$$\nabla_{\mathbf{b}_1} \text{Div} = \nabla_{\mathbf{z}_1} \text{Div}$$

In some problems we will also want to compute the derivative w.r.t. the input

# The Backward Pass

- Set  $\mathbf{y}_N = Y, \mathbf{y}_0 = \mathbf{x}$
- Initialize: Compute  $\nabla_{\mathbf{y}_N} Div = \nabla_Y Div$
- For layer  $k = N$  down to 1:
  - Compute  $J_{\mathbf{y}_k}(\mathbf{z}_k)$ 
    - Will require intermediate values computed in the forward pass
  - Backward recursion step:
$$\nabla_{\mathbf{z}_k} Div = \nabla_{\mathbf{y}_k} Div J_{\mathbf{y}_k}(\mathbf{z}_k)$$
$$\nabla_{\mathbf{y}_{k-1}} Div = \nabla_{\mathbf{z}_k} Div \mathbf{W}_k$$
  - Gradient computation:
$$\nabla_{\mathbf{W}_k} Div = \mathbf{y}_{k-1} \nabla_{\mathbf{z}_k} Div$$
$$\nabla_{\mathbf{b}_k} Div = \nabla_{\mathbf{z}_k} Div$$

# The Backward Pass

- Set  $\mathbf{y}_N = Y, \mathbf{y}_0 = \mathbf{x}$
- Initialize: Compute  $\nabla_{\mathbf{y}_N} Div = \nabla_Y Div$
- For layer  $k = N$  down to 1:
  - Compute  $J_{\mathbf{y}_k}(\mathbf{z}_k)$ 
    - Will require intermediate values computed in the forward pass
  - Backward recursion step: Note analogy to forward pass
$$\nabla_{\mathbf{z}_k} Div = \nabla_{\mathbf{y}_k} Div J_{\mathbf{y}_k}(\mathbf{z}_k)$$
$$\nabla_{\mathbf{y}_{k-1}} Div = \nabla_{\mathbf{z}_k} Div \mathbf{W}_k$$
  - Gradient computation:
$$\nabla_{\mathbf{W}_k} Div = \mathbf{y}_{k-1} \nabla_{\mathbf{z}_k} Div$$
$$\nabla_{\mathbf{b}_k} Div = \nabla_{\mathbf{z}_k} Div$$

# For comparison: The Forward Pass

- Set  $\mathbf{y}_0 = \mathbf{x}$
- For layer  $k = 1$  to  $N$  :

– Forward recursion step:

$$\mathbf{z}_k = \mathbf{W}_k \mathbf{y}_{k-1} + \mathbf{b}_k$$

$$\mathbf{y}_k = f_k(\mathbf{z}_k)$$

- Output:

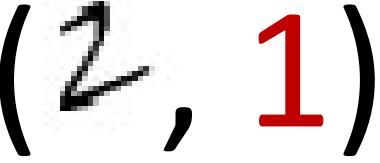
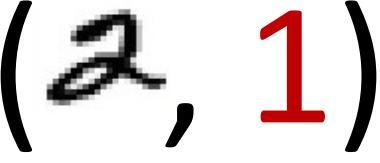
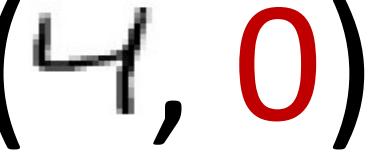
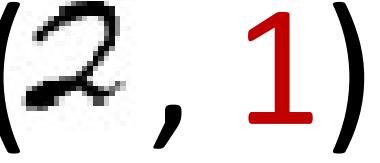
$$\mathbf{Y} = \mathbf{y}_N$$

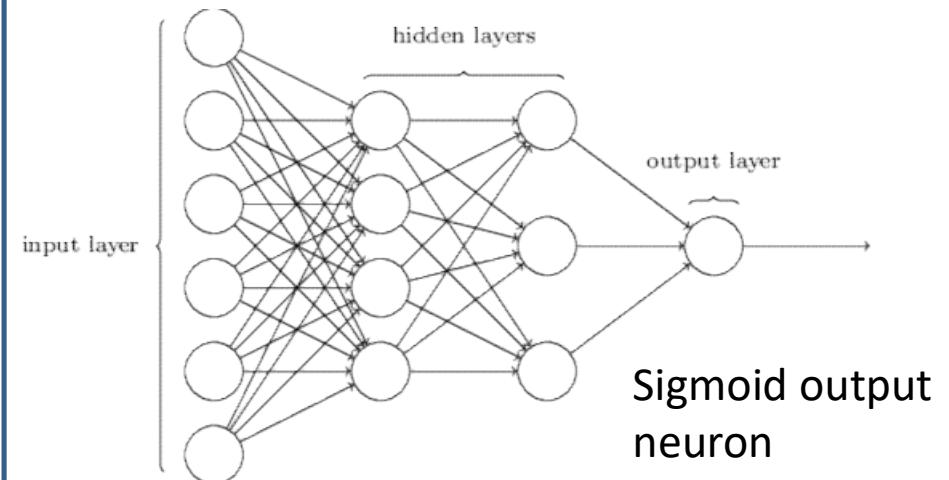
# Neural network training algorithm

- Initialize all weights and biases ( $\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2, \dots, \mathbf{W}_N, \mathbf{b}_N$ )
- Do:
  - $Err = 0$
  - For all  $k$ , initialize  $\nabla_{\mathbf{W}_k} Err = 0, \nabla_{\mathbf{b}_k} Err = 0$
  - For all  $t = 1:T$  # Loop through training instances
    - Forward pass : Compute
      - Output  $\mathbf{Y}(X_t)$
      - Divergence  $Div(\mathbf{Y}_t, \mathbf{d}_t)$
      - $Err += Div(\mathbf{Y}_t, \mathbf{d}_t)$
    - Backward pass: For all  $k$  compute:
      - $\nabla_{\mathbf{y}_k} Div = \nabla_{\mathbf{z}_{k+1}} Div \mathbf{W}_{k+1}$
      - $\nabla_{\mathbf{z}_k} Div = \nabla_{\mathbf{y}_k} Div J_{\mathbf{y}_k}(\mathbf{z}_k)$
      - $\nabla_{\mathbf{W}_k} Div(\mathbf{Y}_t, \mathbf{d}_t) = \mathbf{y}_{k-1} \nabla_{\mathbf{z}_k} Div; \nabla_{\mathbf{b}_k} Div(\mathbf{Y}_t, \mathbf{d}_t) = \nabla_{\mathbf{z}_k} Div$
      - $\nabla_{\mathbf{W}_k} Err += \nabla_{\mathbf{W}_k} Div(\mathbf{Y}_t, \mathbf{d}_t); \nabla_{\mathbf{b}_k} Err += \nabla_{\mathbf{b}_k} Div(\mathbf{Y}_t, \mathbf{d}_t)$
  - For all  $k$ , update:
$$\mathbf{W}_k = \mathbf{W}_k - \frac{\eta}{T} (\nabla_{\mathbf{W}_k} Err)^T; \quad \mathbf{b}_k = \mathbf{b}_k - \frac{\eta}{T} (\nabla_{\mathbf{b}_k} Err)^T$$
- Until  $Err$  has converged

# Setting up for digit recognition

Training data

|   |  |
|---|--|
| (  , 0)  | (  , 1)  |
| (  , 1)  | (  , 0)  |
| (  , 0) | (  , 1) |

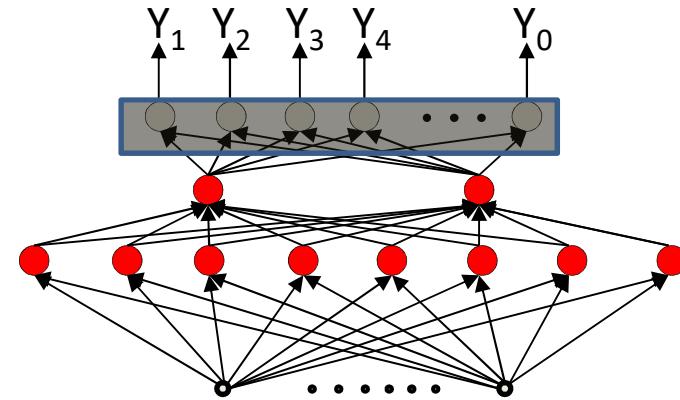


- Simple Problem: Recognizing “2” or “not 2”
- Single output with sigmoid activation
  - $Y \in (0,1)$
  - $d$  is either 0 or 1
- Use KL divergence
- Backpropagation to learn network parameters

# Recognizing the digit

Training data

|        |        |
|--------|--------|
| (Σ, 0) | (2, 1) |
| (2, 1) | (4, 0) |
| (0, 0) | (2, 1) |



- More complex problem: Recognizing digit
- Network with 10 (or 11) outputs
  - First ten outputs correspond to the ten digits
    - Optional 11th is for none of the above
- Softmax output layer:
  - Ideal output: One of the outputs goes to 1, the others go to 0
- Backpropagation with KL divergence to learn network

# Issues

- Convergence: How well does it learn
  - And how can we improve it
- How well will it generalize (outside training data)
- What does the output really mean?
- *Etc..*

# Next up

- Convergence and generalization