

3. Training Deep Networks

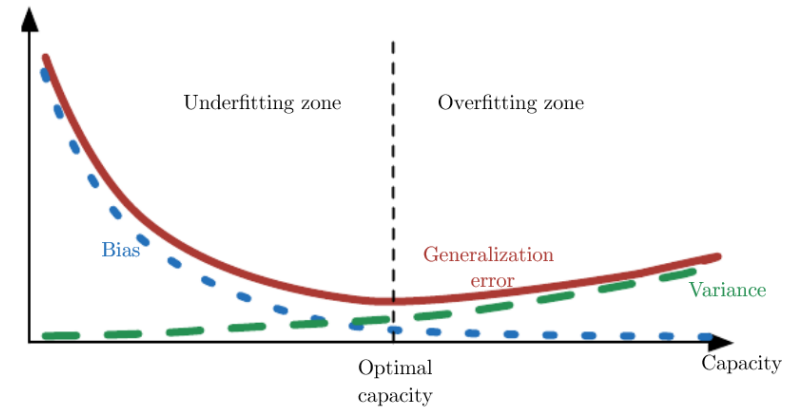
CS 5242 Neural Networks and Deep Learning

YOU, Yang

30.08.2022

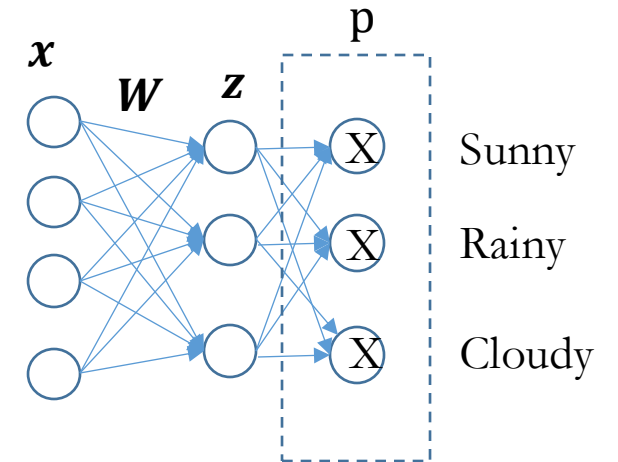
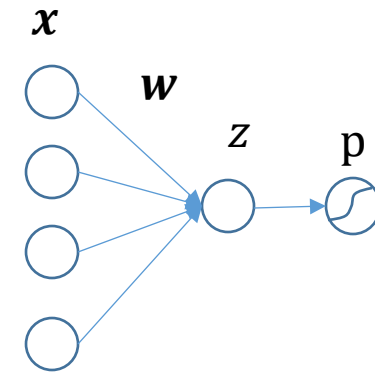
Recap

- Overfitting, underfitting, bias and variance
 - Polynomial regression
 - Hyper-parameter (degree M)
 - Dataset splitting
 - Training, validation, and test data
 - Tune the parameters to minimize training error
 - Overfit training data \rightarrow evaluate on the validation data
 - Tune the hyper-parameters to minimize the validation error
 - Overfit validation data \rightarrow evaluate on the test data



Recap

- Perceptron model
- Binary classification model
 - Logistic function \rightarrow probability
 - Binary cross-entropy loss
- Multi-class classification model
 - Softmax/multinomial regression
 - Softmax function \rightarrow a vector of probabilities
 - Cross-entropy loss



Agenda

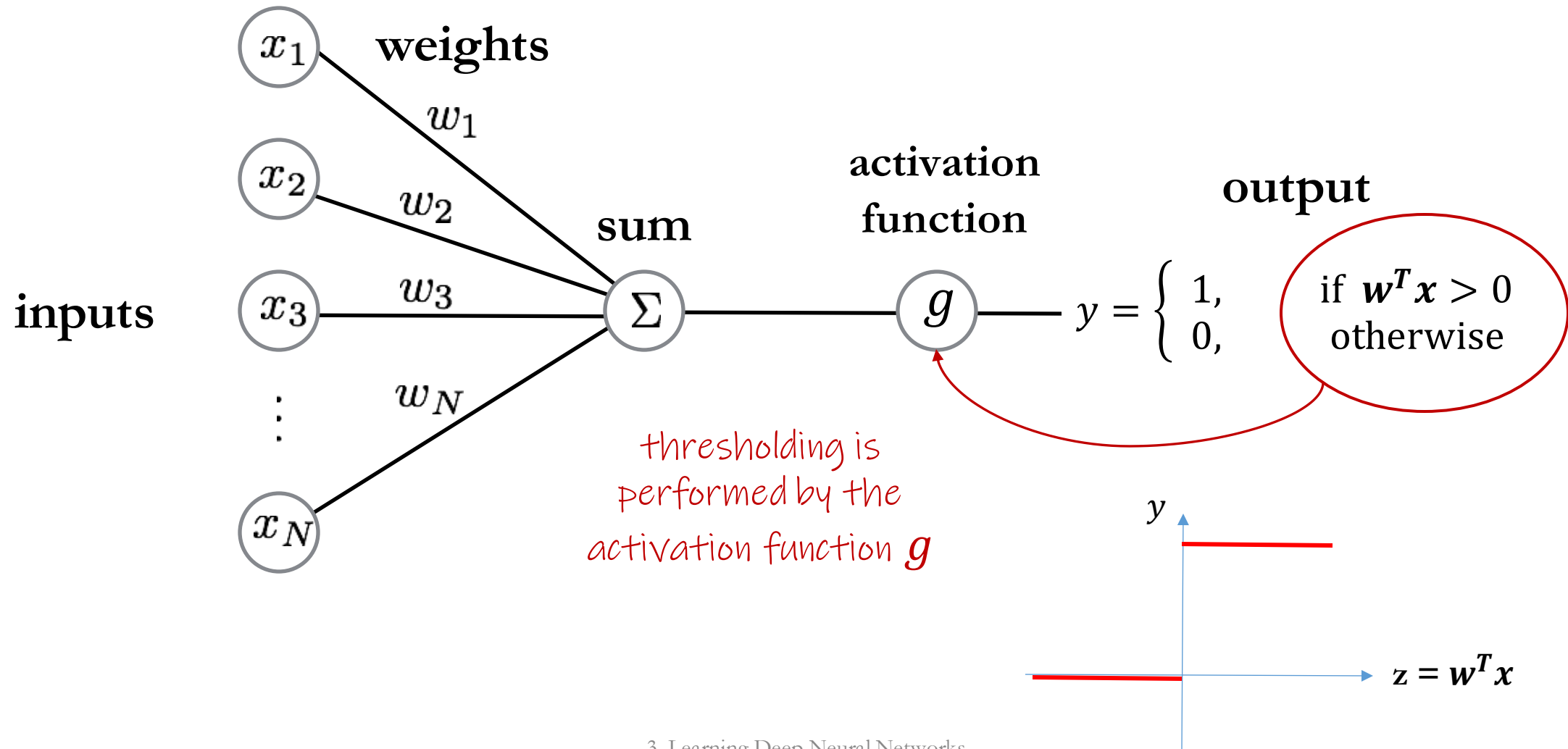
- Multi-layer Perceptrons
 - Activation functions & non-linear feature transformations
 - Going “deeper”
- Backpropagation algorithm
 - For computing the gradients of a deep neural network
- Advanced optimization algorithms
 - Stochastic gradient descent (SGD)
 - SGD+
- Training tricks
 - Data normalization
 - Parameter initialization

Multi-Layer Perceptrons (MLP)

Non-linear Activation Functions

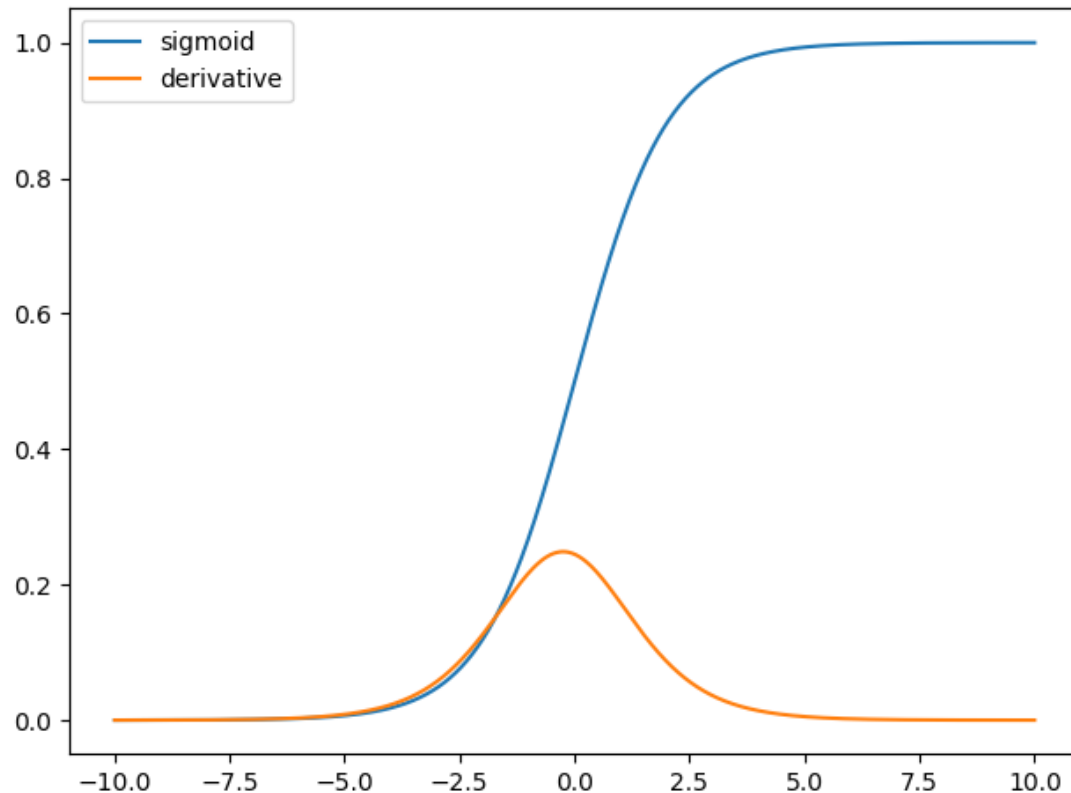
Definition of MLPs

The Perceptron



Logistic / Sigmoid Activation

We've already seen another type of activation function: Logistic or sigmoid function.



<https://i.stack.imgur.com/inMoa.png>

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

$$\frac{\partial \sigma}{\partial z} = \sigma(z)(1 - \sigma(z))$$

detailed derivation [[here](#)]

Activation functions $g()$

A linear function or mapping $f(x)$ has properties of

- **additivity** $f(x + y) = f(x) + f(y)$ and
- **homogeneity** $f(\alpha x) = \alpha f(x)$.

- Activation functions perform non-linear feature transformations

Q: Why do we need non-linear activation functions?

- (1) Linear transformations alone are limited.
- (2) Gives meaning to multi-layer perceptrons, because otherwise, linear layers would collapse (explain in later slides)

name	plot	equation	derivative	range
Binary step		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x \neq 0 \\ ? & \text{for } x = 0 \end{cases}$	$\{0, 1\}$
Logistic (a.k.a. Soft step)		$f(x) = \frac{1}{1 + e^{-x}}$	$f'(x) = f(x)(1 - f(x))$	$(0, 1)$
TanH		$f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1$	$f'(x) = 1 - f(x)^2$	$(-1, 1)$
ArcTan		$f(x) = \tan^{-1}(x)$	$f'(x) = \frac{1}{x^2 + 1}$	$(-\frac{\pi}{2}, \frac{\pi}{2})$
Softsign [7][8]		$f(x) = \frac{x}{1 + x }$	$f'(x) = \frac{1}{(1 + x)^2}$	$(-1, 1)$
Rectified linear unit (ReLU) ^[9]		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$[0, \infty)$
Leaky rectified linear unit (Leaky ReLU) ^[10]		$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0.01 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$(-\infty, \infty)$
Parametric rectified linear unit (PReLU) ^[11]		$f(\alpha, x) = \begin{cases} \alpha x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(\alpha, x) = \begin{cases} \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$(-\infty, \infty)$
Randomized leaky rectified linear unit (RRReLU) ^[12]		$f(\alpha, x) = \begin{cases} \alpha x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$ [1]	$f'(\alpha, x) = \begin{cases} \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$(-\infty, \infty)$

Source from: https://en.wikipedia.org/wiki/Activation_function

Linear functions are limited

- Simple binary example:

$$y = \begin{cases} 1, & \text{if } \mathbf{w}^T \mathbf{x} > 0 \\ 0, & \text{otherwise} \end{cases}$$

“Training” algorithm

For each training example (\mathbf{x}, y)

Compute the prediction \tilde{y}

Update $w_i = w_i + \alpha * (y - \tilde{y}) * x_i$

Possible weights for AND function:

$$w_1 = 1, w_2 = 1, b = -1.5$$

- **Multi-layer perceptrons with only linear functions are still single-layer perceptrons.**
- **Perceptrons are mainly termed as “linear classifiers” and can be used only for linear separable applications.**
- **XOR is not linearly separable.**
- **We need real multi-layer perceptrons.**

AND function

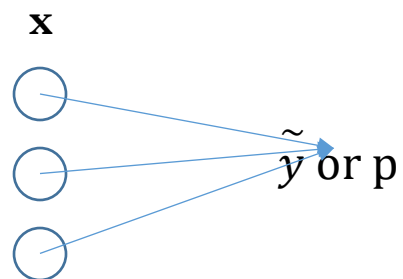
x1	x2	y
1	1	1
1	0	0
0	0	0
0	1	0

XOR function

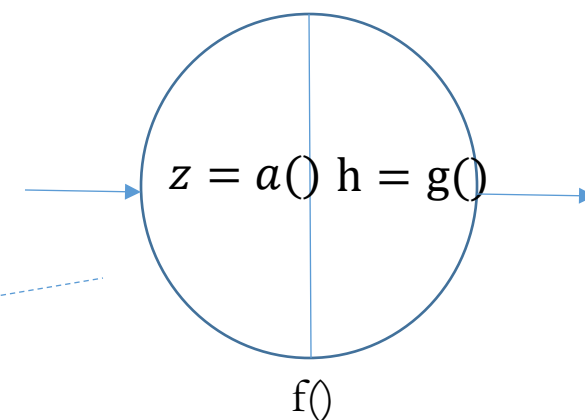
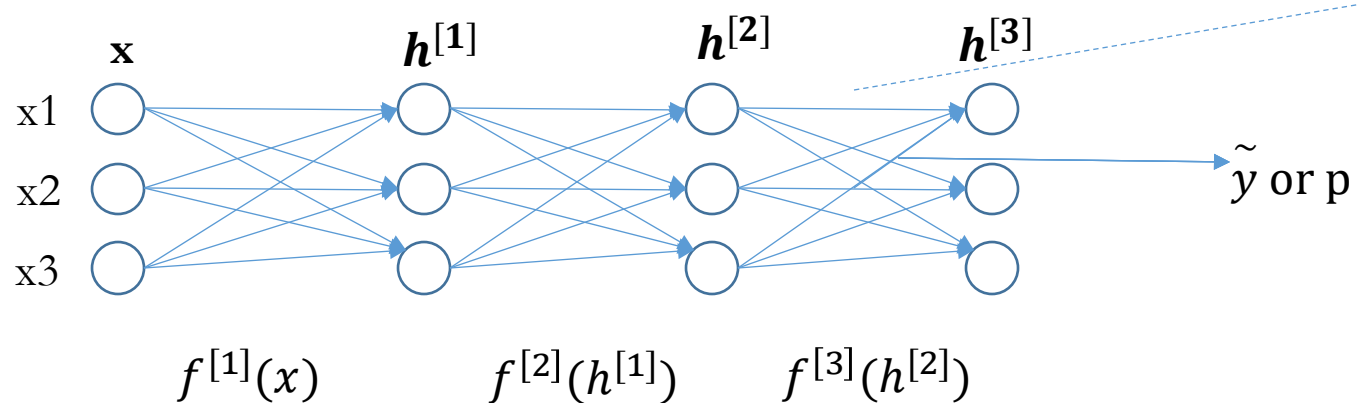
x1	x2	y
0	0	0
0	1	1
1	0	1
1	1	0

Multi-layer Perceptron

single layer
perceptron
(no hidden layers)



multi-layer
perceptron



Multi-layer Perceptron

A net with multiple layers that transform input features into hidden features and then make predictions

- At least one hidden layer
- i^{th} layer consists of a **linear/affine** transformation function

$$\mathbf{z}^{[i]} = a^{[i]}(\mathbf{h}^{[i-1]}) = W^{[i]}\mathbf{h}^{[i-1]} + \mathbf{b}^{[i]}$$

$$W^{[i]} \in R^{n_i \times n_{i-1}}, \mathbf{b}^i \in R^{n_i}$$

- n_i is the number of hidden units at the i^{th} layer and is a tunable hyper-parameter
- followed by a **non-linear activation** function

$$\mathbf{h}^{[i]} = g^{[i]}(\mathbf{z}^{[i]}), \in R^{n_i}$$

A Linear MLP?

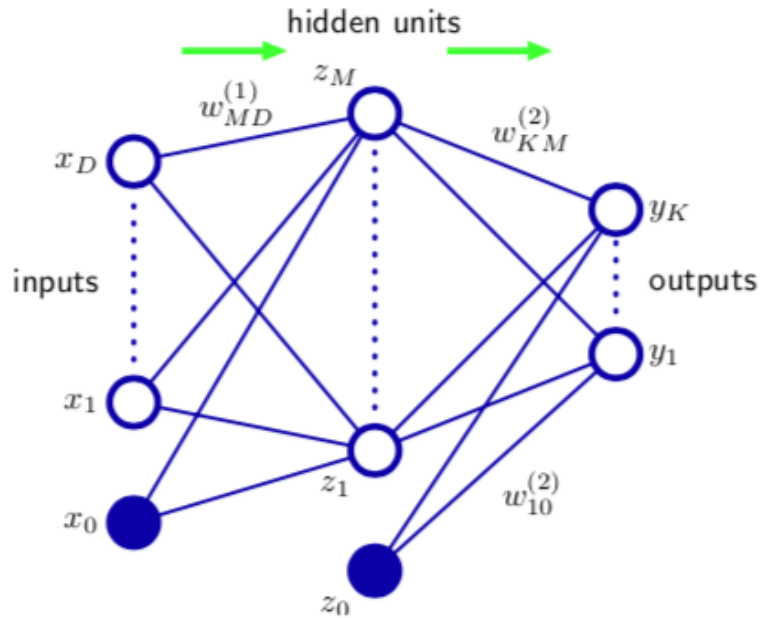


Figure: 5.1 from Bishop

Suppose we have a network with N hidden layers where activations h were linear



$$\begin{aligned}
 y_k(\mathbf{x}, \mathbf{w}) &= \sigma \left(\mathbf{w}_{n\cdot}^{(N)} \cdot h^{(N-1)} \dots h^{(2)} \left(\mathbf{w}_{k\cdot}^{(2)} \cdot h^{(1)} \left(\mathbf{w}_{j\cdot}^{(1)} \cdot \mathbf{x} \right) \right) \right) \\
 &= \sigma \left(\mathbf{w}_{n\cdot}^{(N)} \cdot h^{(N-1)} \dots h^{(2)} \left(\mathbf{w}_{k\cdot}^{(2)} \cdot \underbrace{\mathbf{w}_{j\cdot}^{(1)} \cdot h^{(1)}(\mathbf{x})}_{\text{weighting outside of activation}} \right) \right) \\
 &= \sigma \left(\underbrace{\mathbf{w}_{n\cdot}^{(N)} \dots \mathbf{w}_{k\cdot}^{(2)} \cdot \mathbf{w}_{j\cdot}^{(1)}}_{\mathbf{w}'} \cdot \underbrace{h^{(N-1)} \circ h^{(2)} \circ h^{(1)}}_g(\mathbf{x}) \right) \\
 &= \sigma(\mathbf{w}' g(\mathbf{x}))
 \end{aligned}$$

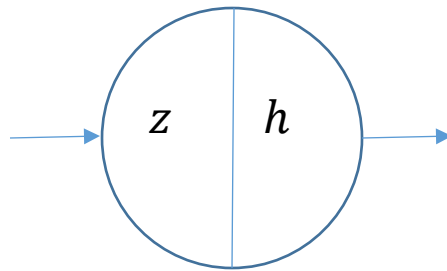
This result says that if all activation functions were linear, we can always define an equivalent collapsed network.

Because successive linear transformations together form yet another linear transformation, a multi-layered linear network is not so interesting. Therefore, we need non-linear activations.

Activation functions

- Logistic (Sigmoid, σ) VS ReLU

Logistic (a.k.a. Soft step)		$f(x) = \frac{1}{1 + e^{-x}}$	$f'(x) = f(x)(1 - f(x))$	$(0, 1)$
Rectified linear unit (ReLU) ^[9]		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$[0, \infty)$



$$h^{[i]} = g^{[i]}(z^{[i]})$$

Logistic activation

If z_k is large, e.g. >10 , then h_k is near 1

If z_k is small, e.g. <-10 , Then h is near 0

For both cases, $\frac{\partial h_k}{\partial z_k} \approx 0 \rightarrow$ gradient vanishing
 \rightarrow gradient vanishing

ReLU activation

If z_k is positive, $\frac{\partial h_k}{\partial z_k} = 1$, no gradient vanishing

If z_k is negative, $\frac{\partial h_k}{\partial z_k} = 0$ no gradients \cup

better than sigmoid, since z_k has a larger working domain
Leaky ReLU (see previous slides) resolves the gradient vanishing problem for negative z_k

Training MLP

- Cross-entropy loss for classification
- Squared Euclidean distance for regression Squared L2 Norm
- GD algorithm
 - Compute $J(\mathbf{X}, \mathbf{Y})$
 - Compute gradient : $\frac{\partial J}{\partial \mathbf{W}^{[1]}}, \frac{\partial J}{\partial \mathbf{b}^{[1]}}, \frac{\partial J}{\partial \mathbf{W}^{[2]}}, \frac{\partial J}{\partial \mathbf{b}^{[2]}} \dots$
 - Update: $\mathbf{W}^{[k]} = \mathbf{W}^{[k]} - \alpha \frac{\partial J}{\partial \mathbf{W}^{[k]}}$, $\mathbf{b}^{[k]} = \mathbf{b}^{[k]} - \alpha \frac{\partial J}{\partial \mathbf{b}^{[k]}}$

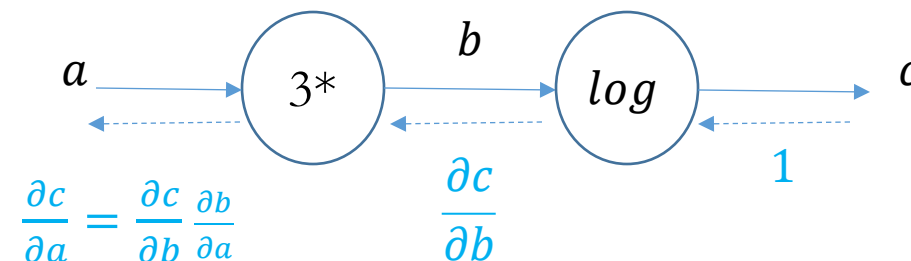
Backpropagation

How to compute the gradient?

- GD algorithm needs the gradient of the loss w.r.t each parameter
- Linear regression model
 - $\frac{\partial L}{\partial \mathbf{w}} = (\tilde{y} - y)\mathbf{x}$
- Logistic regression model
 - $\frac{\partial L}{\partial \mathbf{w}} = (p - y)\mathbf{x}$
- Softmax regression model
 - $\frac{\partial L}{\partial \mathbf{w}} = (\mathbf{p} - \mathbf{y})\mathbf{x}^T$
- How about MLP model with 10 hidden layers?
 - We need a **modular** way to compute the gradients

Chain rule I

- $c = \log 3a$ Given $a = 1$, what is $\frac{\partial c}{\partial a}$?



- $b = f_1(a) = 3a,$ $c = f_2(b) = \log b \rightarrow c = \log 3a$
- $\frac{\partial c}{\partial a} = \frac{\partial c}{\partial b} \frac{\partial b}{\partial a} = f_2'(b) f_1'(a) = \frac{1}{b} * 3 = \frac{1}{3a} * 3 = \frac{3}{3} = 1$

Chain rule I

- Basic chain rule

- $v_2 = f_1(v_1), v_3 = f_2(v_2) \quad \dots \quad v_k = f_{k-1}(v_{k-1})$

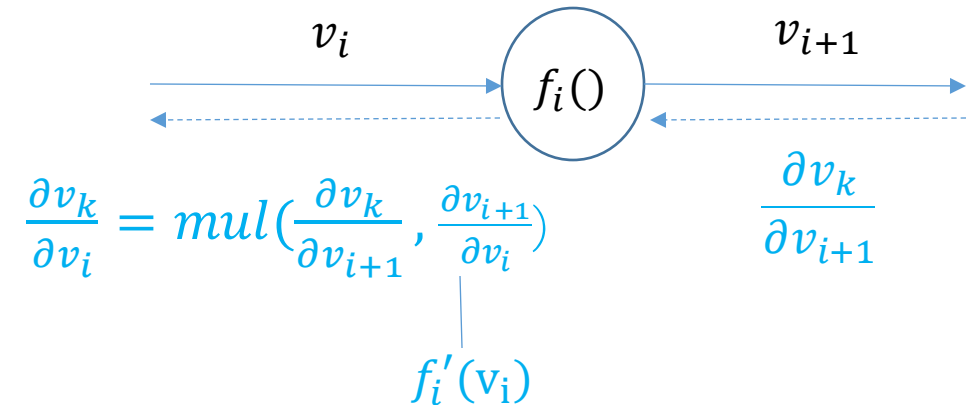
- v_i could be a scalar, vector, matrix, tensor

- $\frac{\partial v_k}{\partial v_i} = \text{mul}\left(\frac{\partial v_k}{\partial v_{i+1}}, \frac{\partial v_{i+1}}{\partial v_i}\right)$

General multiplication operation;
to avoid the scalar/matrix/tensor
transpose/ordering messiness ...

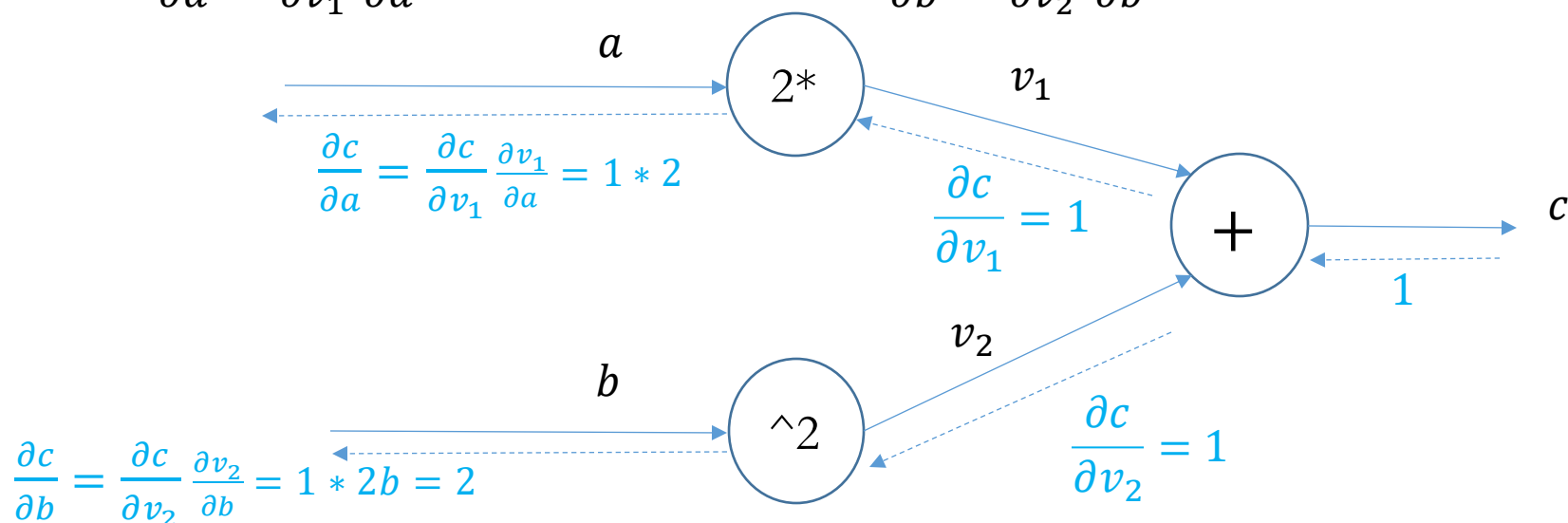
Some
special
cases;
shape-
check to
confirm!

- Reorder and transpose
 - $\frac{\partial v_k}{\partial v_{i+1}} \frac{\partial v_{i+1}}{\partial v_i}$ if all are scalars
 - $\frac{\partial v_{i+1}}{\partial v_i} \frac{\partial v_k}{\partial v_{i+1}}$ if v_k is a scalar, v_{i+1}, v_i are vectors



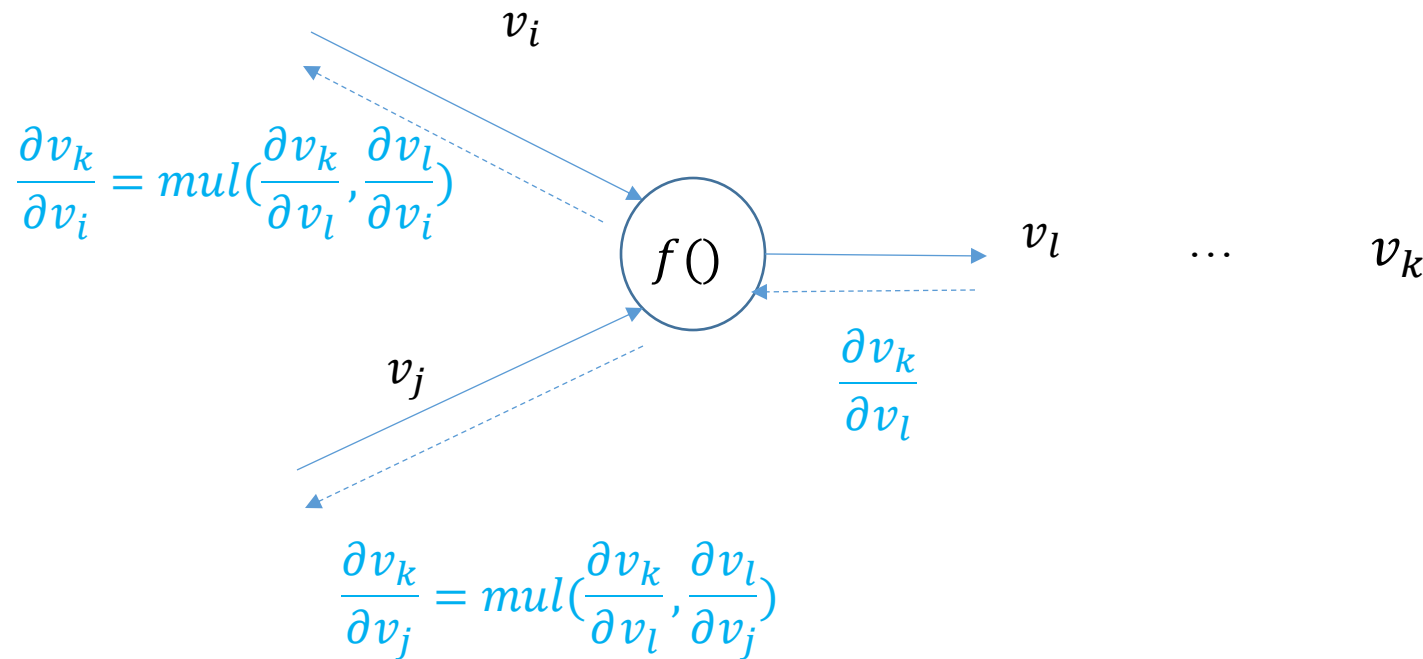
Chain rule II

- $c = f(a, b) = 2a + b^2$
 - Given, $a = 1, b = 1$
 - $\frac{\partial c}{\partial a} = 2, \frac{\partial c}{\partial b} = 2b = 2$
- $v_1 = f_1(a) = 2a, v_2 = f_2(b) = b^2, c = f_3(v_1, v_2) = v_1 + v_2$
 - $\frac{\partial c}{\partial a} = \frac{\partial c}{\partial v_1} \frac{\partial v_1}{\partial a} = 1 * 2 = 2$
 - $\frac{\partial c}{\partial b} = \frac{\partial c}{\partial v_2} \frac{\partial v_2}{\partial b} = 1 * 2b = 2$



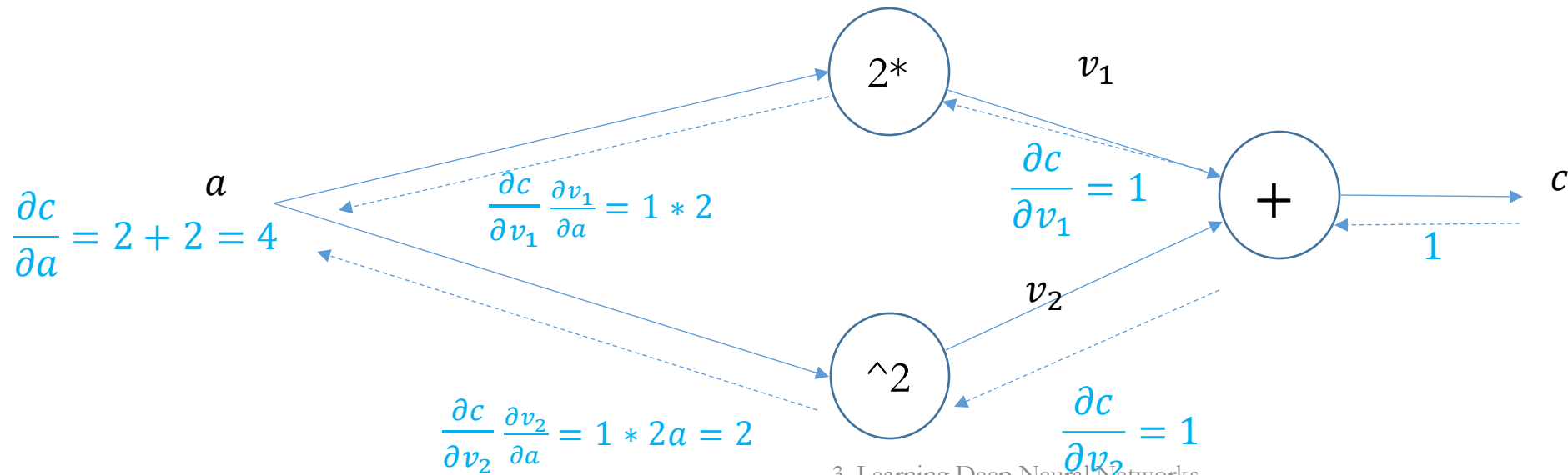
Chain rule II

- $v_l = f(v_i, v_j)$,
- v_k is the final output



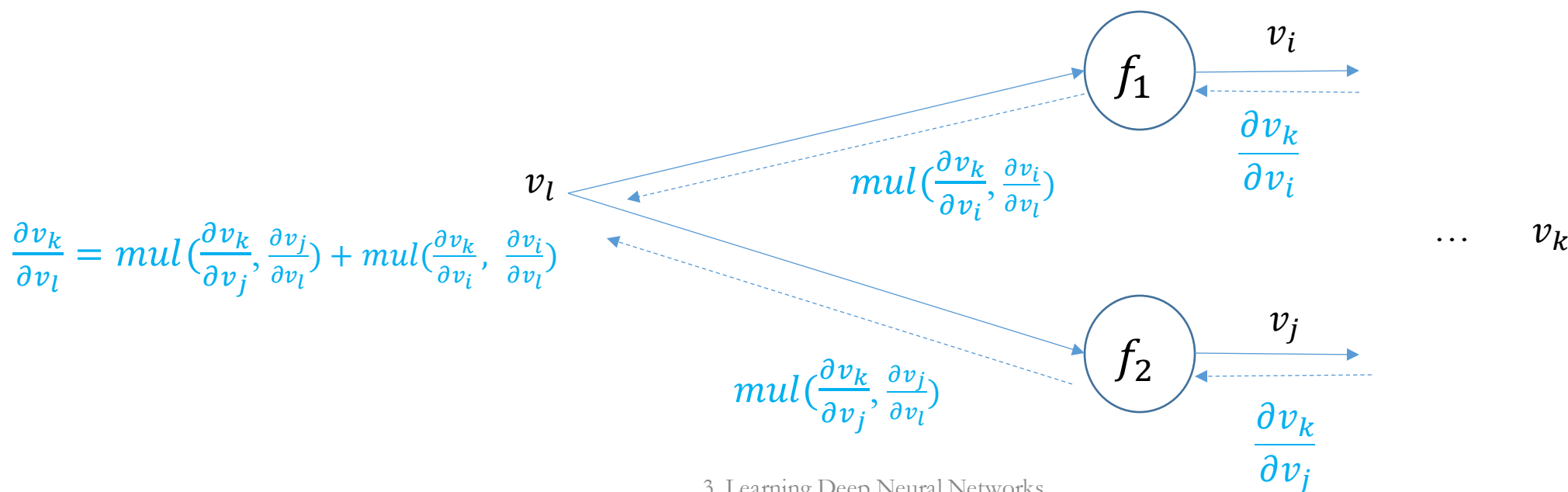
Chain rule III

- $c = 2a + a^2$
 - Given $a = 1$, $\frac{\partial c}{\partial a} = 2 + 2a = 4$
- $v_1 = f_1(a) = 2a, v_2 = f_2(a) = a^2, c = f_3(v_1, v_2) = v_1 + v_2$
 - $\frac{\partial c}{\partial a} = \frac{\partial c}{\partial v_1} \frac{\partial v_1}{\partial a} + \frac{\partial c}{\partial v_2} \frac{\partial v_2}{\partial a} = 1 * 2 + 1 * 2a = 4$



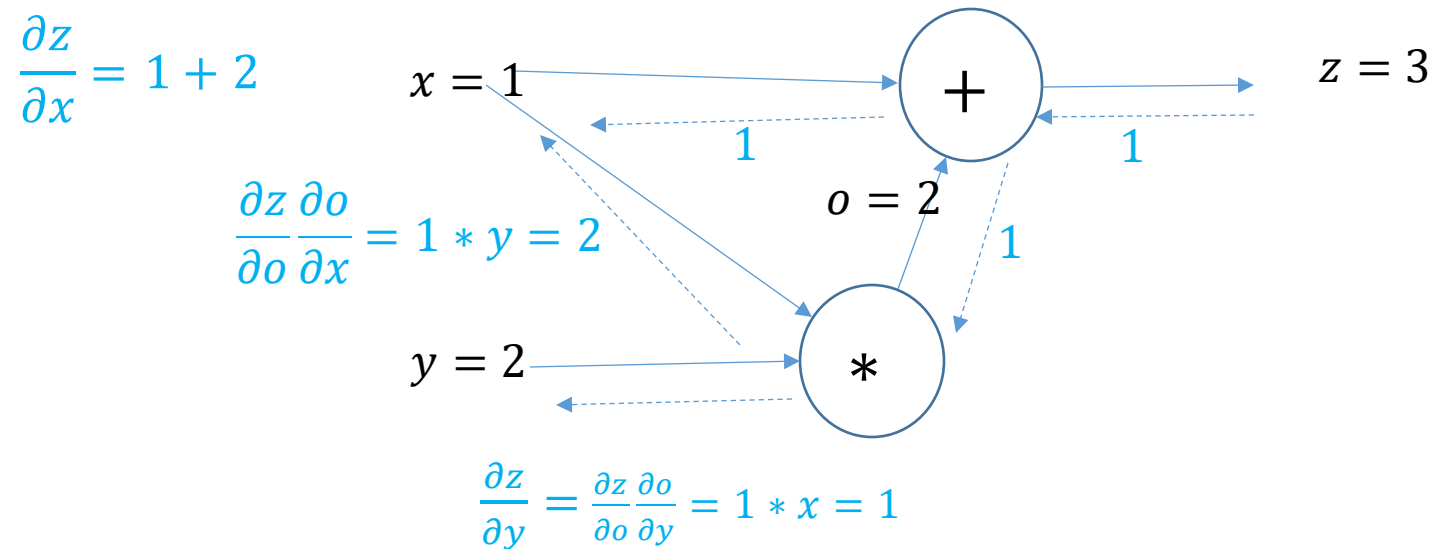
Chain rule III

- $v_i = f_1(v_l)$
- $v_j = f_2(v_l)$
- v_k is the final output



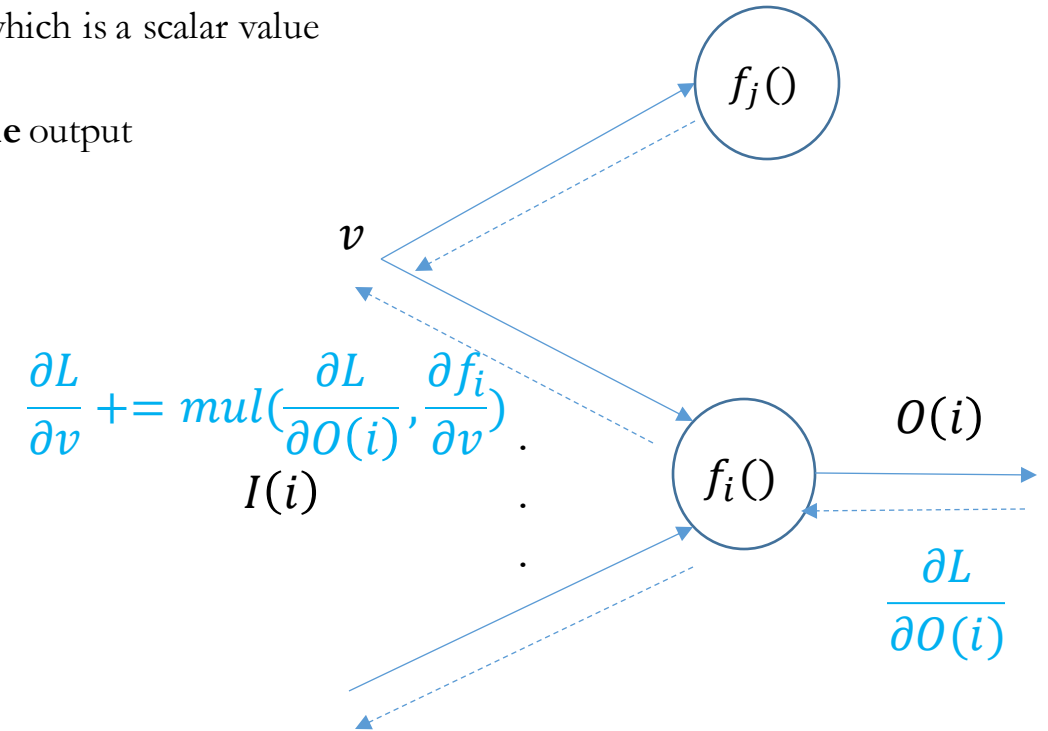
Practice

- $z = f(x, y) = x * y + x$
- Given $x = 1, y = 2$, compute $\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}$?



Backpropagation (BP)

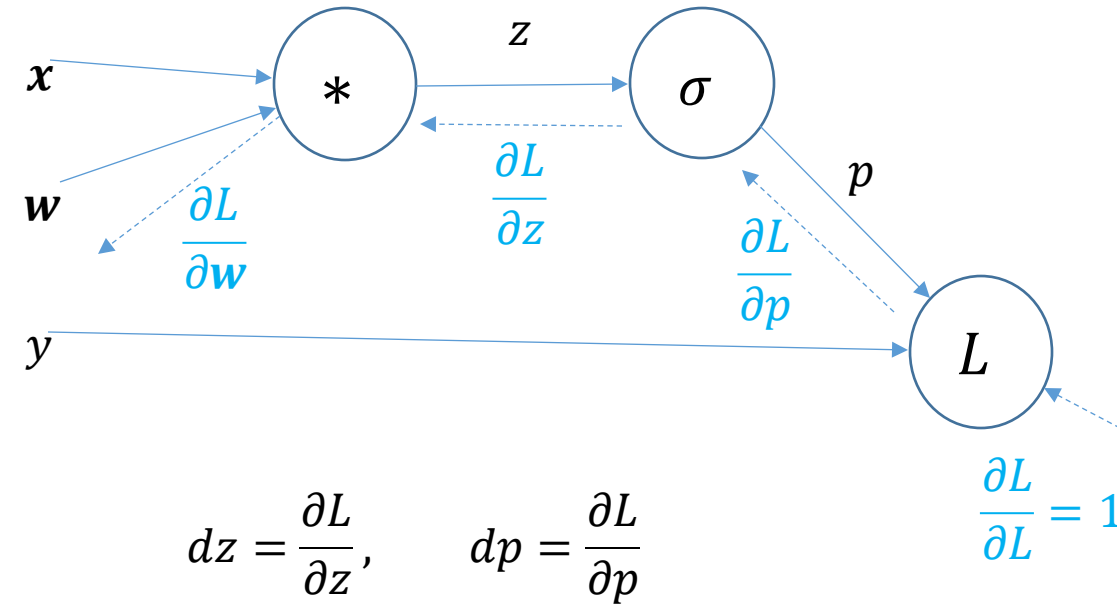
- Computation graph
 - Denote all variables as v_1, v_2, \dots sorted in topological order
 - The last variable is the loss (L) for neural network models, which is a scalar value
 - Denote all operations as f_1, f_2, \dots, f_k
 - Each operation accepts multiple inputs and generate **a single** output
 - which may be used by multiple other operations
 - $I(i)$ denotes all the input variables to f_i
 - $O(i)$ denotes the single output variable of f_i
- Forward pass
 - For $i = 1, 2, \dots, k$
 - Run operation $f_i: I(i) \rightarrow O(i)$
- Backward pass
 - Initialize $\frac{\partial L}{\partial v} = 0$ for all v
 - For $i = k, \dots, 1$ **Note the reverse order!**
 - For each $v \in I(i)$ Compute $\frac{\partial L}{\partial v} += mul(\frac{\partial L}{\partial O(i)}, \frac{\partial f_i(I(i))}{\partial v})$



Logistic regression

- $z = \mathbf{w}^T \mathbf{x}, \mathbf{x} \in R^{n \times 1}, \mathbf{w} \in R^{n \times 1}$
- $p = \sigma(z)$
- $L(p, y) = -y \log p - (1 - y) \log(1 - p)$
- Dot
 - Forward(\mathbf{x}, \mathbf{w}): $\mathbf{w}^T \mathbf{x}$
 - Backward($dz, \mathbf{x}, \mathbf{w}$): $dz \mathbf{x}$
- Logistic
 - Forward(z): $\sigma(z)$
 - Backward(dp, z): $p = \sigma(z); dp * p * (1 - p) = (-\frac{y}{p} + \frac{1-y}{1-p}) * p * (1 - p) = p - y$
- Binary-Cross-entropy
 - Forward(p, y): $-y \log p - (1 - y) \log(1 - p)$
 - Backward($[dL], p, y$): $-\frac{y}{p} + \frac{1-y}{1-p}$

dL is redundant since it is 1

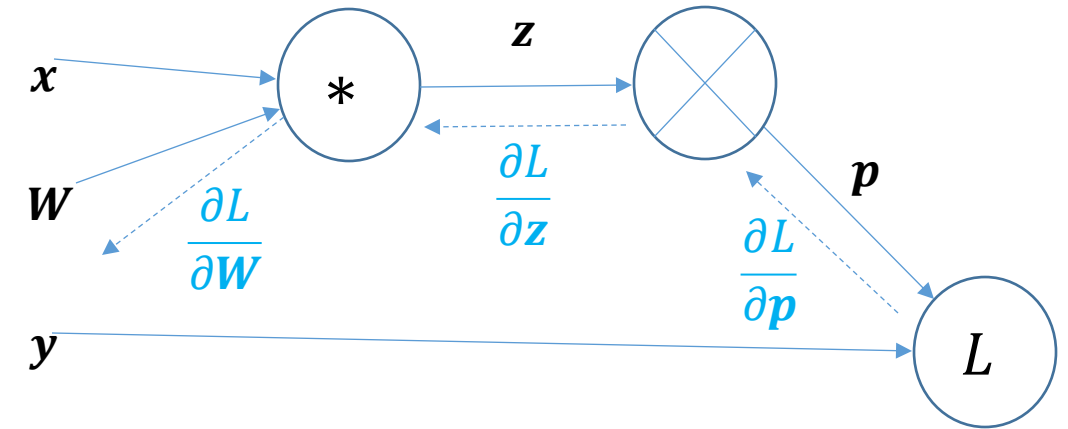


This operation can also be further decomposed; but since binary cross-entropy is used so commonly, we treat it as a whole stand-alone unit.

Softmax regression

- $\mathbf{z} = \mathbf{W}\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^n, \mathbf{W} \in \mathbb{R}^{k \times n}$
- $\mathbf{p} = \text{softmax}(\mathbf{z})$ $\mathbf{p} \in \mathbb{R}^k$
- $L(\mathbf{p}, \mathbf{y}) = \sum_{i=1} -y_i \log p_i$

- Matmul **matrix multiplication**
 - Forward(\mathbf{x}, \mathbf{W}): $\mathbf{W}\mathbf{x}$
 - Backward($d\mathbf{z}, \mathbf{x}, \mathbf{W}$)
- Softmax
 - Forward(\mathbf{z}): \mathbf{p}
 - Backward($d\mathbf{p}, \mathbf{z}$)
- Cross-entropy
 - Forward(\mathbf{p}, \mathbf{y}): $\sum_{i=1} -y_i \log p_i$
 - Backward(\mathbf{p}, \mathbf{y})

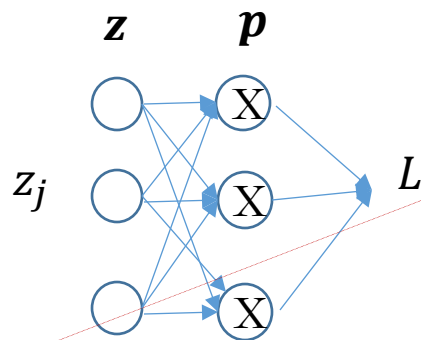


Cleaner and computationally cheaper to consider the softmax and cross-entropy together!

$$\frac{\partial L}{\partial \mathbf{z}} = \frac{\partial L}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \mathbf{z}} = \mathbf{p} - \mathbf{y}$$

easy messy matrix, since num & denom are vectors

$$\begin{aligned} \frac{\partial L}{\partial z_j} &= \sum_i \frac{\partial L}{\partial p_i} \frac{\partial p_i}{\partial z_j} \\ &= \frac{\partial L}{\partial p_j} \frac{\partial p_j}{\partial z_j} + \sum_{i \neq j} \frac{\partial L}{\partial p_i} \frac{\partial p_i}{\partial z_j} \\ &= -\frac{y_j}{p_j} (p_j - p_j^2) + \sum_{i \neq j} -\frac{y_i}{p_i} (-p_i p_j) \\ &= -y_j(1 - p_j) + \sum_{i \neq j} y_i p_j \\ &= -y_j + y_j p_j + p_j \sum_{i \neq j} y_i \quad (\sum y_i = 1) \\ &= -y_j + y_j p_j + p_j(1 - y_j) = p_j - y_j \end{aligned}$$



$$\frac{\partial L}{\partial p_i} = -\frac{y_i}{p_i}, \quad \frac{\partial L}{\partial \mathbf{p}} = -\frac{\mathbf{y}}{\mathbf{p}}$$

$$p_i = \frac{e^{z_i}}{\sum_k e^{z_k}}$$

$$h(x) = \frac{f(x)}{g(x)} \rightarrow h'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)}$$

$$\frac{\partial \sum_k e^{z_k}}{\partial e^{z_j}} = \frac{\partial e^{z_j}}{\partial e^{z_j}} = e^{z_j}$$

$$\frac{\partial p_i}{\partial z_j} = \begin{cases} i = j, \frac{e^{z_j} \sum_k e^{z_k} - e^{z_j} e^{z_j}}{(\sum_k e^{z_k})^2} = p_j - p_j^2 \\ i \neq j, \frac{0 \sum_k e^{z_k} - e^{z_i} e^{z_j}}{(\sum_k e^{z_k})^2} = -p_i p_j \end{cases}$$

Refer to this [website](#) for the derivation process

Softmax regression

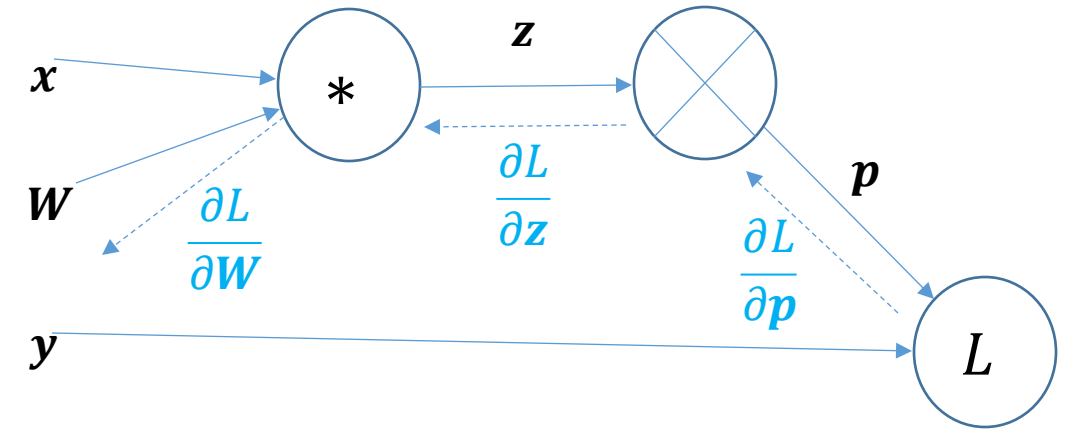
- $\mathbf{z} = \mathbf{W}\mathbf{x}$, $\mathbf{x} \in R^n, \mathbf{W} \in R^{k \times n}$
- $\mathbf{p} = \text{softmax}(\mathbf{z})$ $\mathbf{p} \in R^k$
- $L(\mathbf{p}, \mathbf{y}) = \sum_{i=1} -y_i \log p_i$

- Matmul

- Forward(\mathbf{x}, \mathbf{W}): $\mathbf{W}\mathbf{x}$
- Backward($d\mathbf{z}, \mathbf{x}, \mathbf{W}$): $d\mathbf{z} \mathbf{x}^T$

- Softmax-Cross-Entropy

- Forward(\mathbf{z}): \mathbf{p} ; $\sum_{i=1} -y_i \log p_i$
- Backward(\mathbf{z}, \mathbf{y}): $\mathbf{p} - \mathbf{y}$

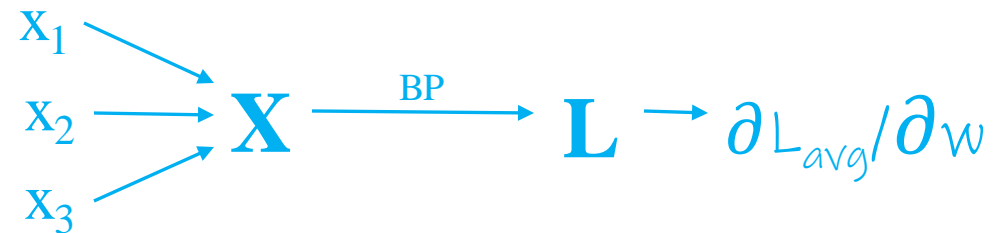
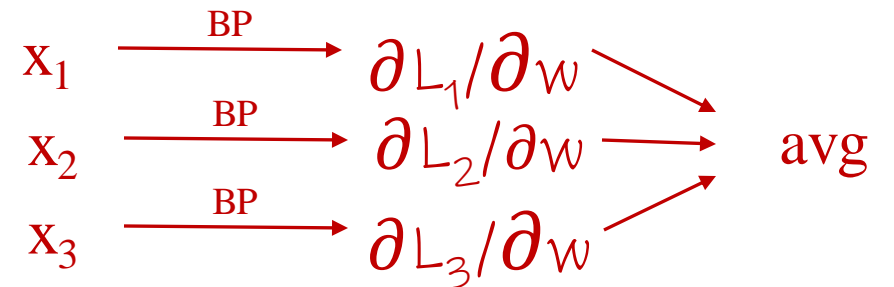


L is a complex node

BP for Multiple examples

So far, we've considered backpropagation for a single example.
How should we handle all of our training samples?

- Approach I (individual)
 - Backpropagation for each example separately
 - Average the gradients across all examples
- Approach II (vectorized)
 - Use matrix (one row per example) in the BP
 - Average the loss
 - Compute gradient based on averaged loss



Which approach is better and why?

BP operations

These are all derived via chain rule

- Add_bias e.g. $Z = XW + b$
 - A is a matrix; b is a row vector
 - Forward(A, b): $C = A + b$
 - Backward(dC, A, b): $dA = dC, db = \mathbf{1}^T dC$
- Array and scalar multiplication
 - v is an array, k is a scalar
 - Forward(v, k): $c = k v$
 - Backward(dc, v, k): $dv = k dc$

BP operations

- Matmul matrix multiplication operation
 - $\mathbf{A} \in R^{m \times k}, \mathbf{B} \in R^{k \times n}$ (including matrix with a single column or row)
 - Forward(\mathbf{A}, \mathbf{B}): $\mathbf{C} = \mathbf{AB} \in R^{m \times n}$
 - Backward($d\mathbf{C}, \mathbf{A}, \mathbf{B}$): $d\mathbf{A} = d\mathbf{C} \mathbf{B}^T, d\mathbf{B} = \mathbf{A}^T d\mathbf{C}$
- Logistic operation
 - \mathbf{a} is an array of any shape
 - Forward(\mathbf{a}): $\mathbf{b} = \sigma(\mathbf{a})$
 - Backward($d\mathbf{b}, \mathbf{a}$): $d\mathbf{a} = d\mathbf{b} * \sigma(\mathbf{a}) * (1 - \sigma(\mathbf{a}))$
- Softmax-Cross-entropy operation
 - \mathbf{Z} and $\mathbf{Y} \in R^{m \times k}$ are matrix of the same shape; each row of \mathbf{a} (or \mathbf{b}) sums to 1
 - Forward (\mathbf{Z}, \mathbf{Y}): $\mathbf{P} = \text{softmax}(\mathbf{Z}); L = \text{sum}(-\mathbf{Y} \log \mathbf{P})/m$
 - Backward(\mathbf{Z}, \mathbf{Y}): $d\mathbf{Z} = (\mathbf{P} - \mathbf{Y})/m$

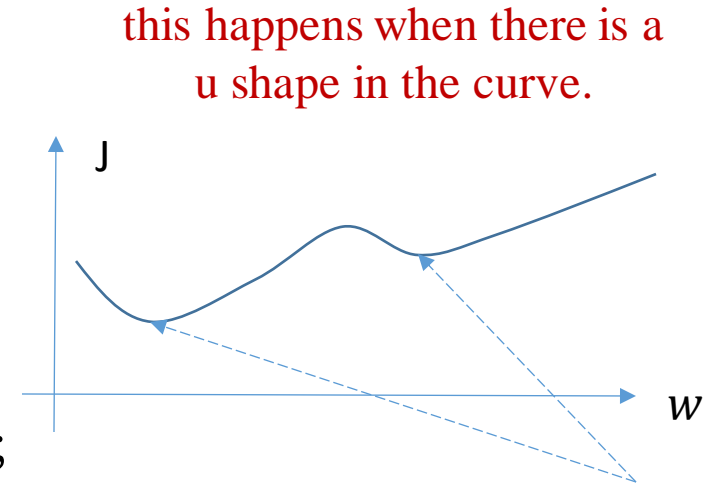
Stochastic Gradient Descent

Gradient descent algorithm

- Define the network architecture
- Select a proper loss function
- Initialize all model parameters (usually randomly)
- Repeat
 - Forward to compute the average loss across all training examples
 - Backward to compute the gradient of the loss w.r.t each parameter
 - $\frac{\partial J}{\partial w}, w \in \Phi$; Φ denotes all parameters of the model
 - Update: $w = w - \alpha \frac{\partial J}{\partial w}, w \in \Phi$

Problem of GD: Local optimum

- For any $w \in \Phi$, $\frac{\partial J}{\partial w} = 0$, local minimum
- This case happens very rarely since:
 - Gradient should be 0
 - Local minimum / u-shape
 - Requires this to be true for all w in the network;



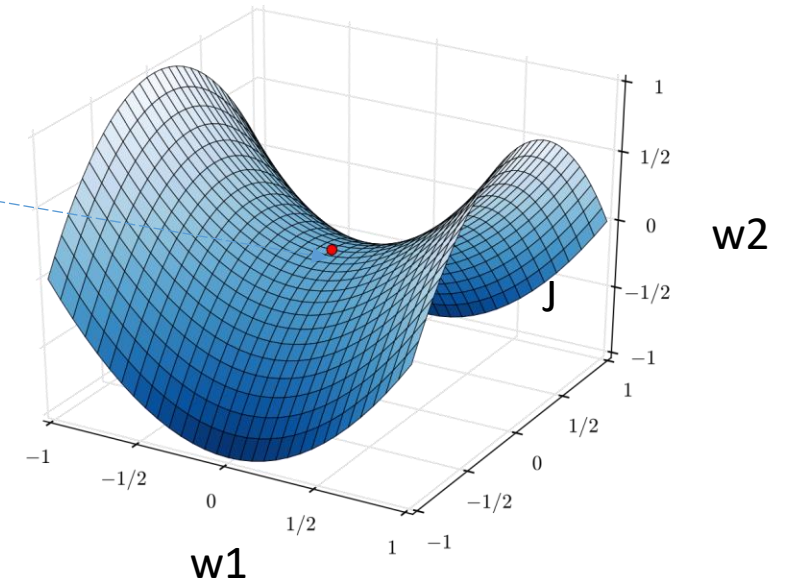
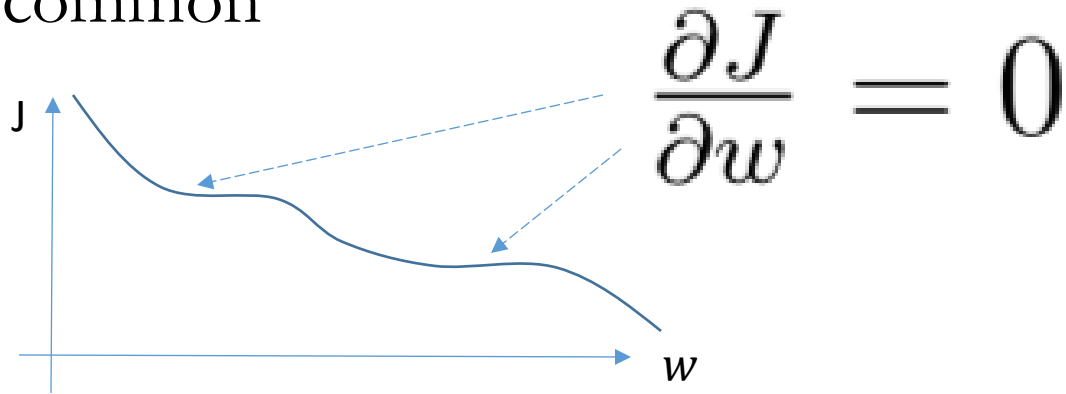
$$w = w - \alpha \frac{\partial J}{\partial w} = w - 0$$

get stuck at local optimum

The gradients are zero, we can't update the weights, so the next iteration's gradients are still zero.
Each iteration is using the same samples: the whole dataset.

Problem of GD: Saddle points [\[link\]](#)

- For any $w \in \Phi$, $\frac{\partial J}{\partial w} = 0$, but is not a local optimum
- More common



These two problems second problem highlights the difficulty of applying gradient descent to non-convex functions. We simply **don't** find the global optimum. Sometimes, we may not even arrive at a local optimum.

from: https://en.wikipedia.org/wiki/Saddle_point

Problem of GD: Efficiency

- GD has to load all training samples into memory to do BP
 - High memory cost
 - High computation cost per iteration

GPT/LLaMA may use 1 trillion tokens, that is around 1 billion samples assuming the sequence length is 1K.
Even for the tiny version of ImageNet (1K classes), the number of samples is 1.3 million.

We can resolve this by making the gradient descent *stochastic*.

An A100 GPU only has 80GB memory.

Stochastic gradient descent (SGD)

- Randomly initialize the parameters
- Repeat until converge
 - Randomly pick a (single) training sample
 - Compute the loss $J^{(i)}$
 - Compute the derivative of $J^{(i)}$ w.r.t each parameter w
 - Update all parameters $w = w - \alpha \frac{\partial J^{(i)}}{\partial w}$
- Advantage
 - More efficient in terms of memory and speed per iteration
 - $\frac{\partial J^{(i)}}{\partial w}$ may not be zero although $\frac{\partial J}{\partial w}=0$, $\frac{\partial J^{(i)}}{\partial w}$ may not be zero although $\frac{\partial J^{(i-1)}}{\partial w}=0$
- Disadvantage
 - Not moving in the optimal direction for every step due to the stochastic process

SGD solves our problem with local optimum and saddle points. How?

We are effectively minimizing the loss for that single example.

Mini-batch SGD

- Randomly initialize the parameters
- Repeat until converge
 - Randomly pick **b** training samples (called a mini-batch)
 - Compute the loss $J = \sum_{i=1}^b J^i / b$
 - Compute the derivative of J w.r.t each parameter w
 - Update each parameter $w = w - \alpha \frac{\partial J}{\partial w}$
- The derivative direction is more stable than SGD
 - Averaged over b samples \rightarrow converges faster in terms of # iterations
- Efficiency
 - R for convergence rate (in proportion to inverse of total number iterations)
 - T for wall time per iteration (e.g. hours)

e.g. **b = 256, 512, 1024**

$$\begin{aligned} R_{GD} &> R_{mini-SGD} > R_{SGD} \\ T_{GD} &> T_{mini-SGD} > T_{SGD} \end{aligned}$$

Batch Size: Small or Big?

- larger batches are more accurate at estimating the gradient, but returns are less than linear
 - In the beginning: 2X batch size, 2X fewer iterations
 - After that: 2X batch size, kX fewer iterations ($k < 2$)
- multicore architectures are underutilized by extremely small batches with some absolute minimum batch size, below which there is a fixed time cost
- if all batch samples are processed in parallel (as is the case for GPUs), then the amount of memory scales w/ batch size, so one is limited by GPU memory
- some hardware have better runtime with specific array sizes, e.g. for GPUs, powers of 2 are usually optimal
- small batches act as regularizers with noisy approximated gradients;
- training with a small batch size may require smaller learning rates to maintain stability due to the high variance in the estimated gradients; this trade-off may slow down the overall learning speed

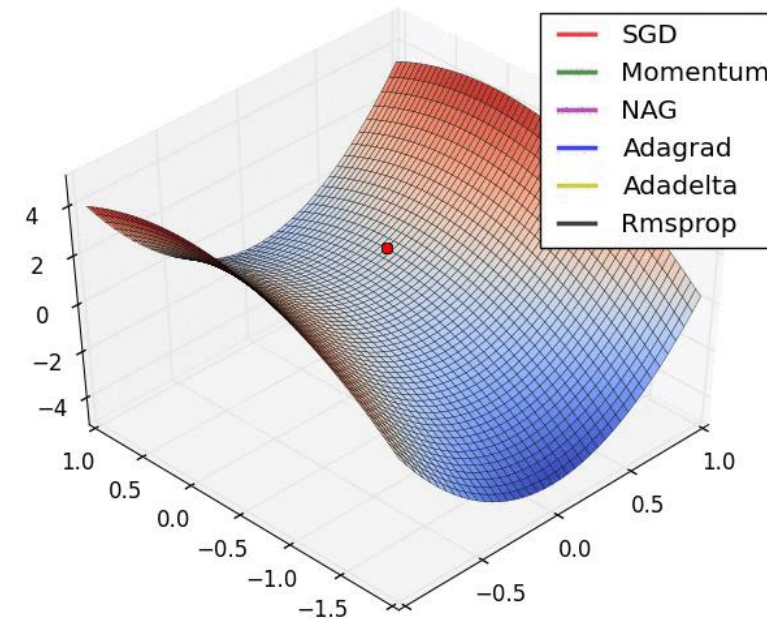
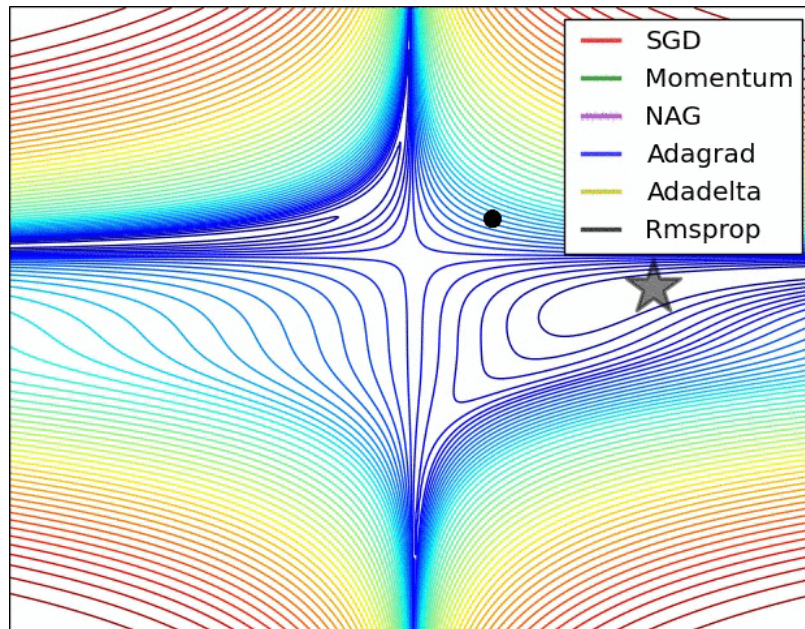
Batch size too small: SGD

Batch size too large: GD

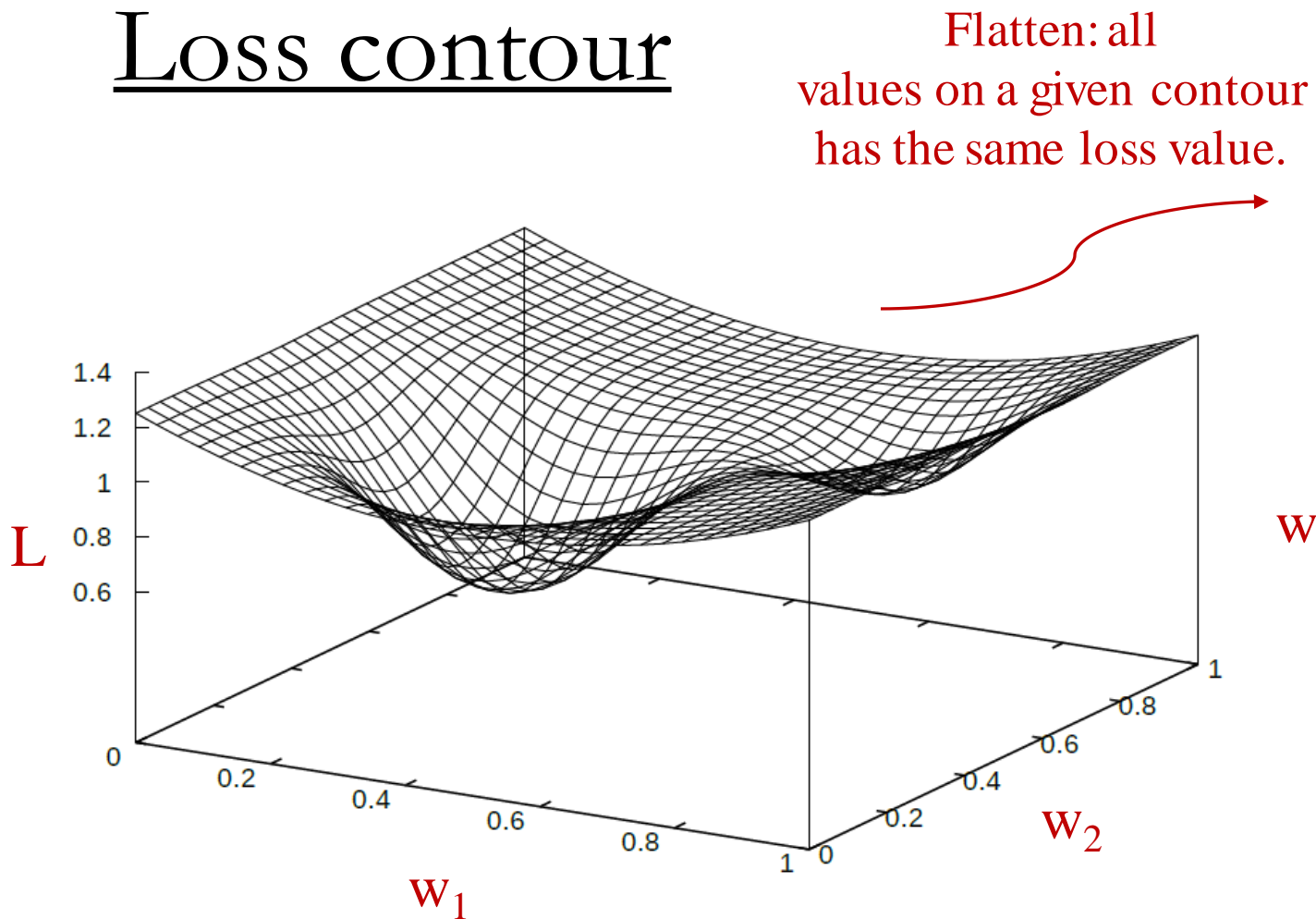
The Evolution of Gradient Descent

- Many variants of gradient descent [\[link\]](https://www.ruder.io/optimizing-gradient-descent/)

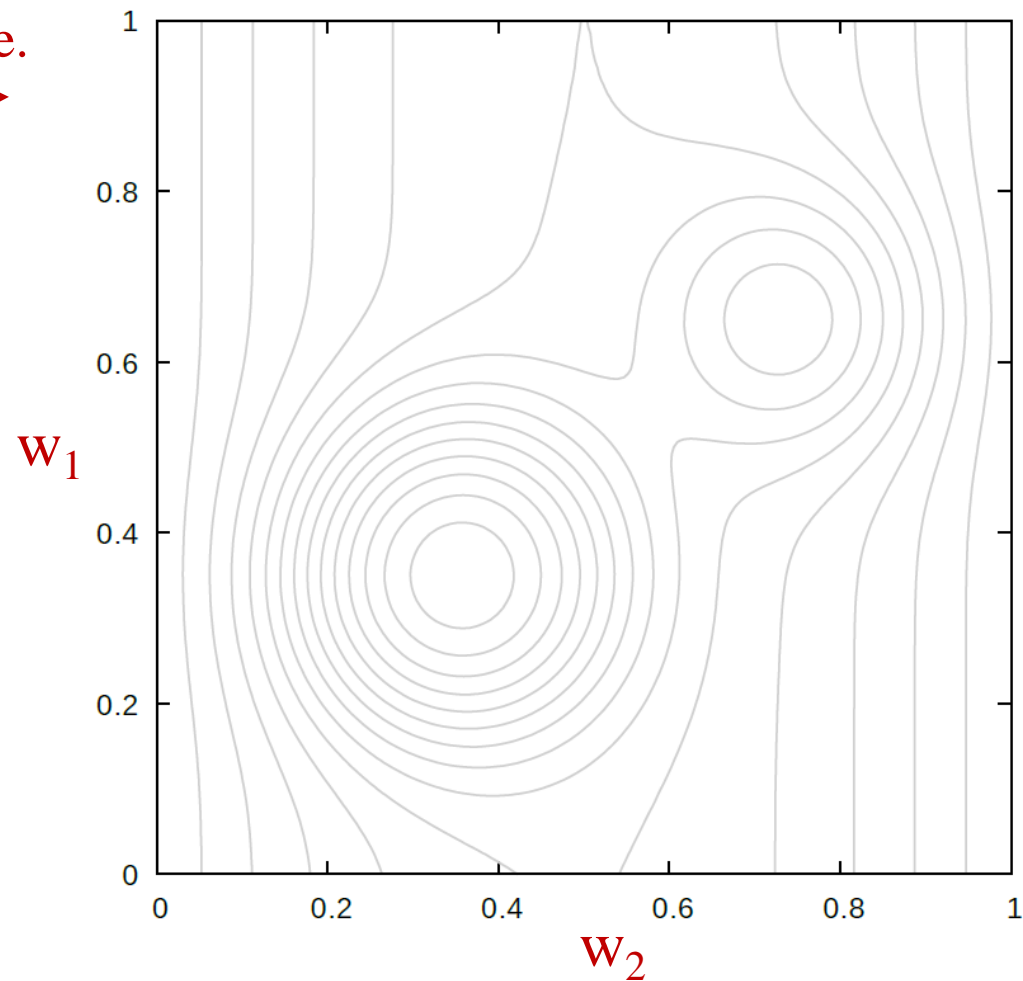
<https://www.ruder.io/optimizing-gradient-descent/>



Loss contour



Consider the loss wrt 2 weights.



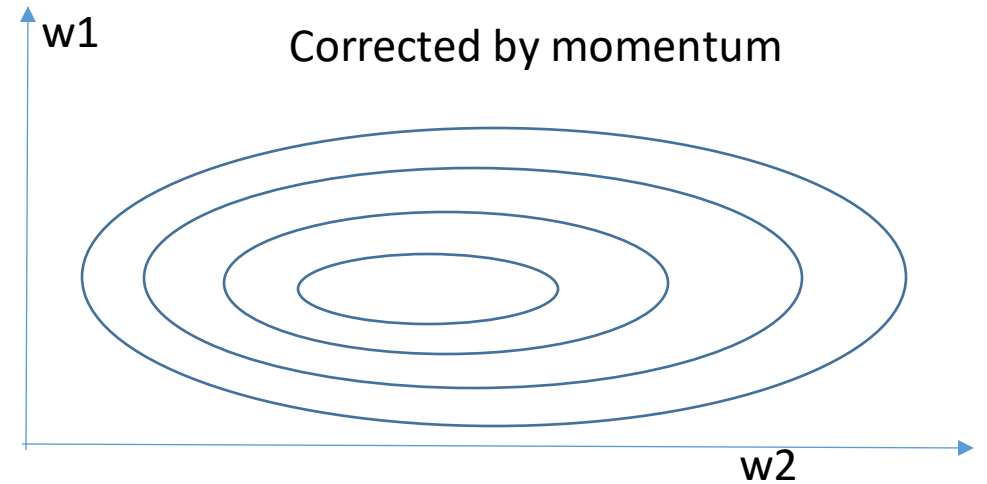
Images from: <https://fleuret.org/ee559/>

3. Learning Deep Neural Networks

Mini-batch SGD with momentum

- Use history movement to correct or accelerate current movement
- Exponential average
 - For every parameter w , we maintain a v for it (initialized as 0)
 - $v = \beta v + g$ (e.g., $\beta = 0.9$)
 - $w = w - \alpha v$

accumulation



We use both the history info and current info to navigate the next step's direction.

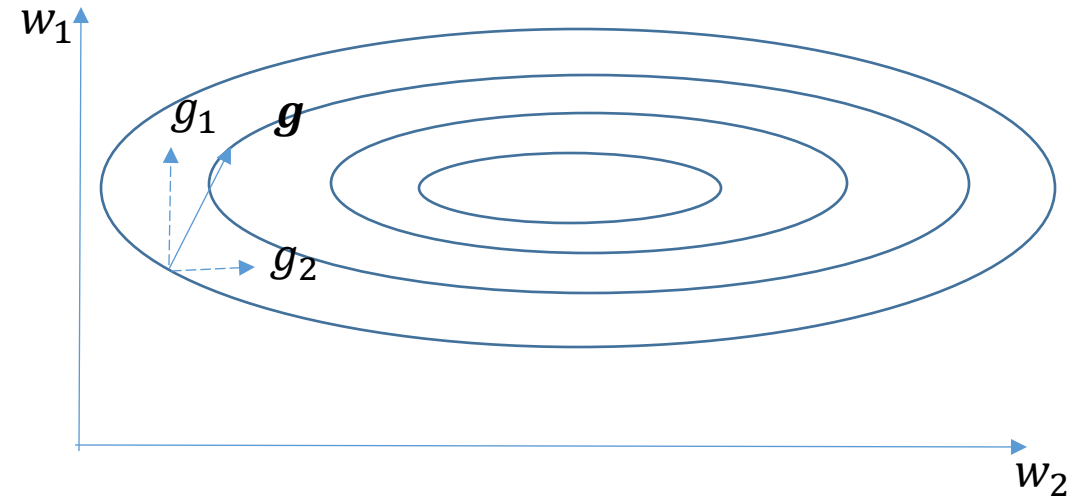
v has all the history information

g only has the current information

We use beta to balance them

RMSprop

- Problem
 - Moving along g_2 would decrease the loss faster, but g_2 is smaller than g_1
- Solution
 - For each parameter W , maintain a variable S (initialized as 0) to rescale the gradient g
 - $s = \beta s + (1 - \beta)g^2$ ($\beta = 0.9$)
 - $w = w - \alpha g / \sqrt{s + \epsilon}$



moving average

s has all the history information
 g only has the current information
We use beta to balance them

Rescales learning rate to remove effect of gradient “size”
Makes a balance between different directions

Adam

Adam offers fast convergence, but sometimes not to good solutions [[link](#)]

- Combines momentum and RMSProp

- $v = \beta_1 v + (1 - \beta_1)g$ (e.g. $\beta_1=0.9$)

moving average of 1st order gradient

- $s = \beta_2 s + (1 - \beta_2)g^2$ (e.g. $\beta_2=0.999$)

moving average of 2nd order gradient

- $w = w - \alpha \frac{\hat{v}}{\sqrt{\hat{s} + \epsilon}}$

- \hat{v} and \hat{s} are bias-corrected for tth iteration

- $\hat{v} = \frac{v}{1 - \beta_1^t}$, $\hat{s} = \frac{s}{1 - \beta_2^t}$

- mainly for the first few iterations;

when t gets large, the denominator goes to 1

$(0.9)^{100} = 0.00002656139$

If $\beta_1 = 0.9$,

$$v_1 = \beta_1 v_0 + (1 - \beta_1)g_1$$

$$= 0.9v_0 + 0.1g_1$$

$$= 0 + 0.1g_1 = 0.1g_1$$

$g_1 = 0.2$, then $v_1 = 0.02$

$$\hat{v} = 0.02 / (1 - 0.9) = 0.02 / 0.1 = 0.2$$

Adam

Algorithm 1: *Adam*, our proposed algorithm for stochastic optimization. See section 2 for details, and for a slightly more efficient (but less clear) order of computation. g_t^2 indicates the elementwise square $g_t \odot g_t$. Good default settings for the tested machine learning problems are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. All operations on vectors are element-wise. With β_1^t and β_2^t we denote β_1 and β_2 to the power t .

Require: α : Stepsize

Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates

Require: $f(\theta)$: Stochastic objective function with parameters θ

Require: θ_0 : Initial parameter vector

$m_0 \leftarrow 0$ (Initialize 1st moment vector)

$v_0 \leftarrow 0$ (Initialize 2nd moment vector)

$t \leftarrow 0$ (Initialize timestep)

while θ_t not converged **do**

$t \leftarrow t + 1$

$g_t \leftarrow \nabla_{\theta} f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)

$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)

$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)

$\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)

$\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)

$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$ (Update parameters)

end while

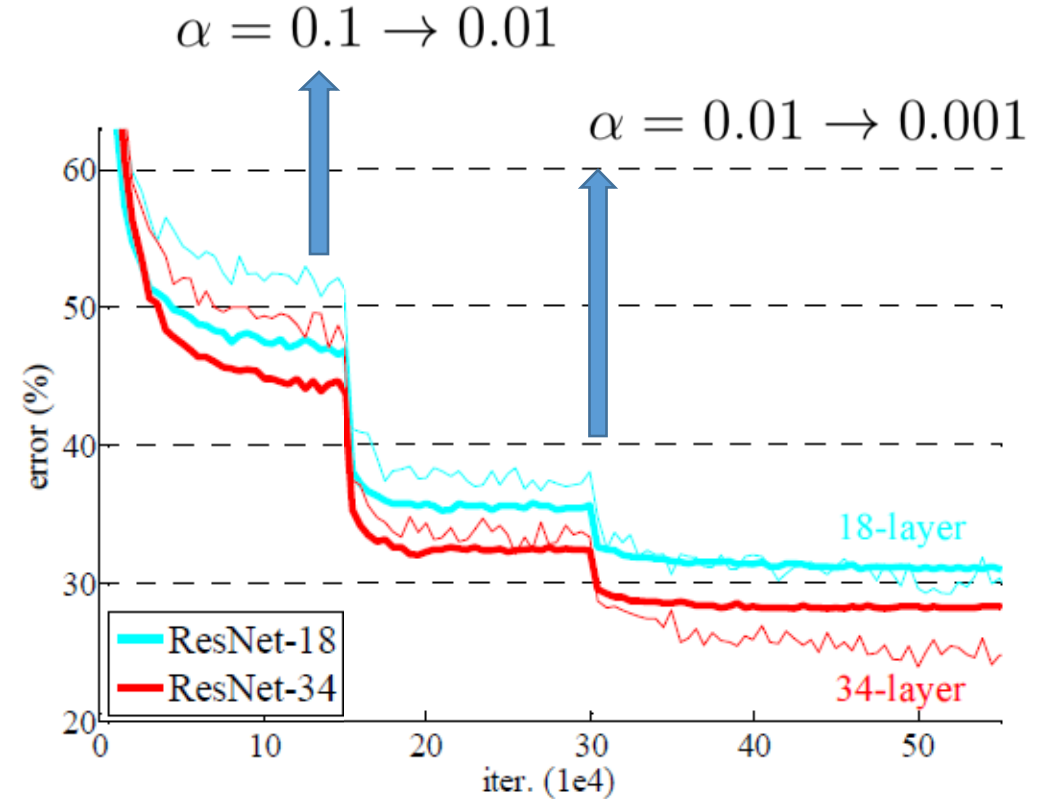
return θ_t (Resulting parameters)

Screenshot from Adam paper: <https://arxiv.org/pdf/1412.6980.pdf>

Learning rate

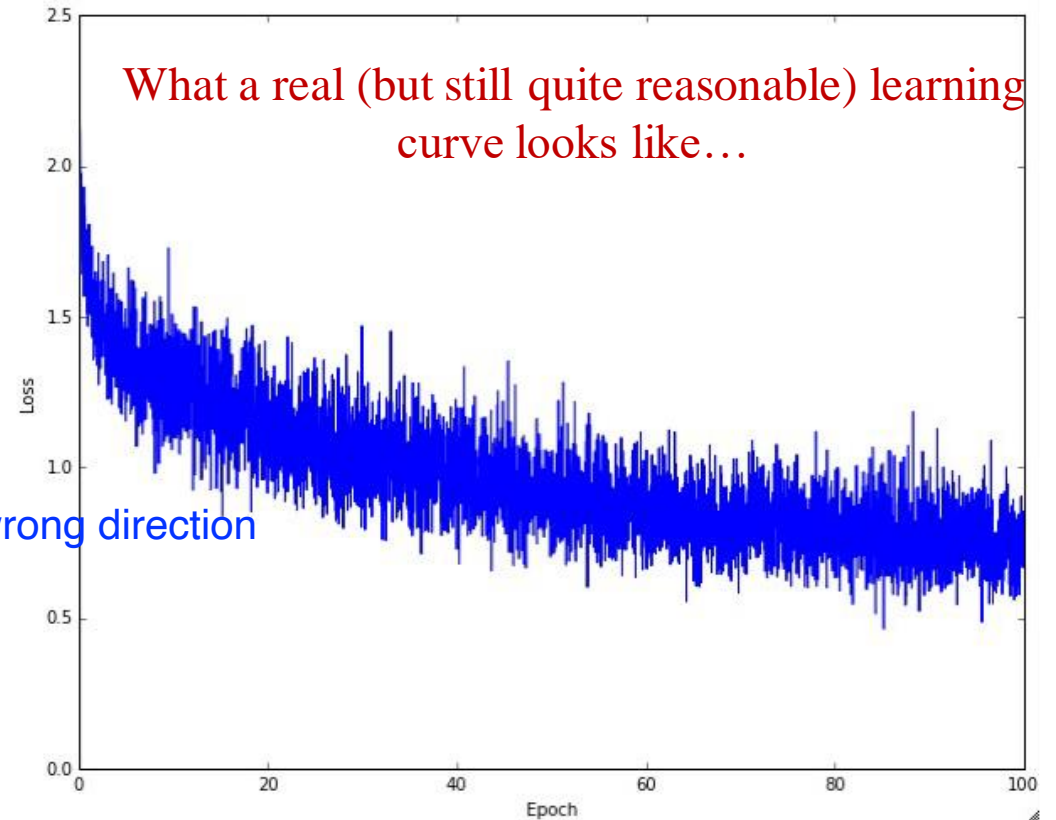
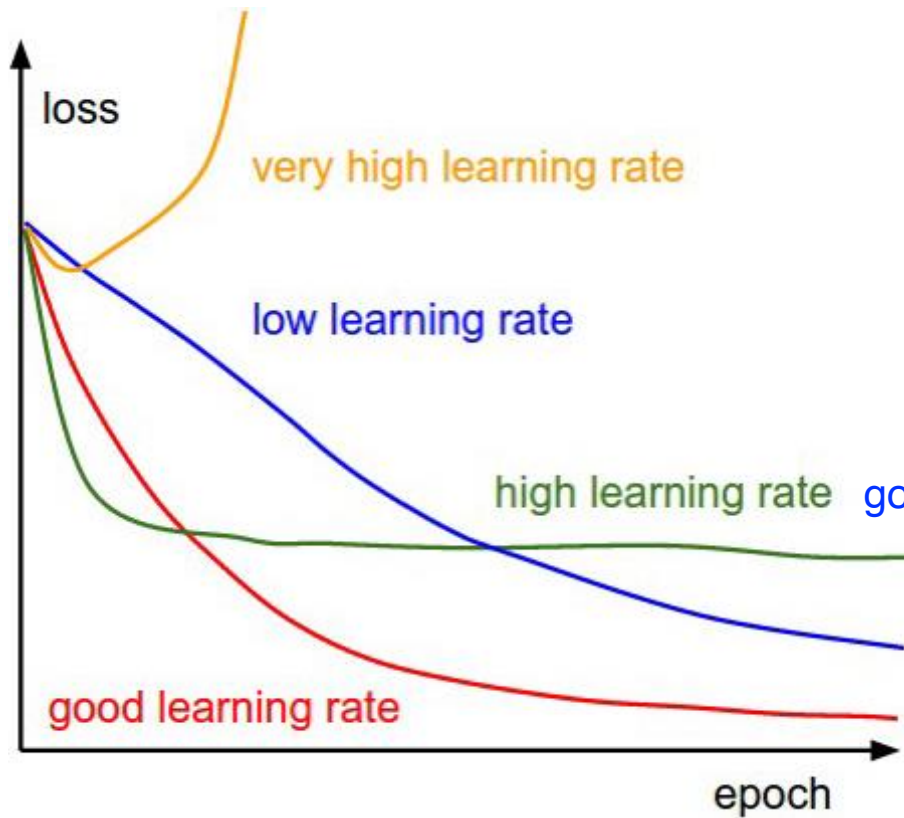
We want to decay the learning rate as we are approaching to convergence. Otherwise, we may jump out of the minimum region.

- Typical starting values: 0.1, 0.01
- Decay learning rate
 - Step
 - $1/t$
 - Exponential, αe^{-kt}
- important for convergence!
 - If it is too large, w will oscillate around the optimal position.
 - If it is too small, it would take many iterations to reach the optimal position.
 - Initialize it with a large value and then decrease it gradually



From [2]

Impact of Learning Rate



Why does the loss increase with a very high learning rate?

The Evolution of Gradient Descent



Tricks of the Trade

Random initialization, data normalization

Regularization (an early look).

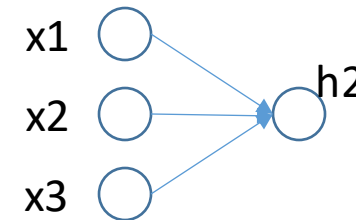
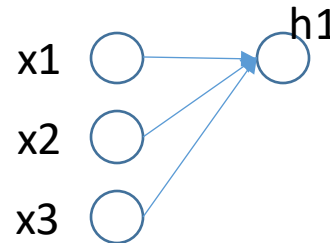
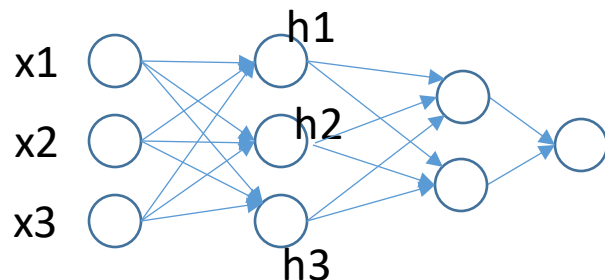
Random parameter initialization

? problems

Why is random initialization so important?

- Suppose that we had all elements of W were initialized as the same value (e.g. 0 or 1)
 - \rightarrow all hidden units are the same
 - \rightarrow derivatives of all hidden units are the same
 - \rightarrow derivatives of weights connecting to x_i are the same
 - \rightarrow weights connecting to x_i have the same update

1. If all neurons in one layer are the same, then only one neuron is enough and all others are redundant \rightarrow a very simple model
2. W 's elements are always the same \rightarrow redundant parameters



Random Initializations Breaks Symmetry

- Initial parameters should break symmetry between different units.
- Two hidden units with the same activation function connected to the same inputs must have different initial parameters.
- If they were the same, then a deterministic learning algorithm, applied to a deterministic cost and model will update both units in the same way.
- If we want each unit to compute a different function (which we do!), then we should use random initialization.

<https://towardsdatascience.com/neural-network-breaking-the-symmetry-e04f963395dd>

How to Randomly Initialize?

- Weight matrix (W)
 - Gaussian, $N(0, 0.01)$
 - Uniform, $U(-0.05, 0.05)$
 - [Glorot/Xavier](#)
 - Gaussian $N(0, \sqrt{2/(\text{fan_in} + \text{fan_out})})$
 - [He/MSRA](#)
 - Gaussian $N(0, \sqrt{2/\text{fan_in}})$
- Too small variance: z vanishing
- Too large variance: z exploding


$$z = XW + b, W \in \mathbb{R}^{n \times k}$$

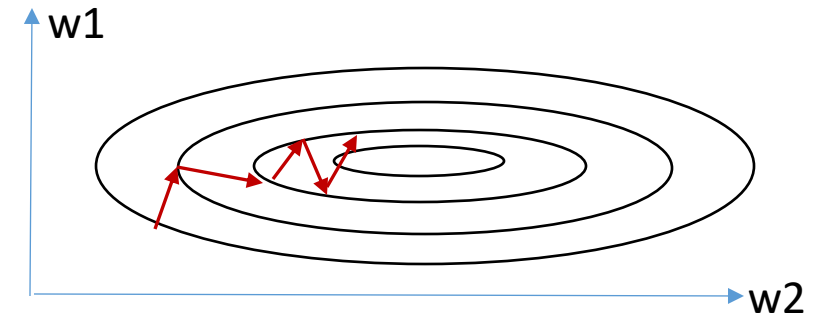
n : fan_in k : fan_out

General idea is to keep the variance of the random values within some reasonable range.

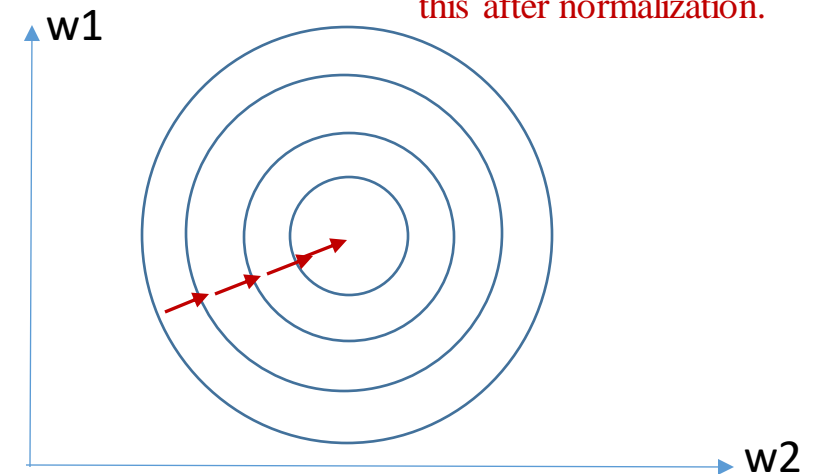
Detailed explanation given [\[here\]](#)

Data normalization

- The features are at difference scale
 - E.g., attributes of a house varies in scale
 - The contours are ellipses
 - The gradient descent algorithm moves along the opposite of the gradient direction
 - **Not the optimal direction**
- Normalize the attributes into similar scale
 - $[-1, 1], [0, 1]$
 - $x_i = \frac{x_i - u_i}{\sigma_i}, x_i = \frac{x_i}{\max(x_i)}$
 - u_i, σ_i, \max are computed over i-th feature of all samples from the training dataset



** this is just a conceptual visualization; it doesn't mean your loss contour looks like this after normalization.

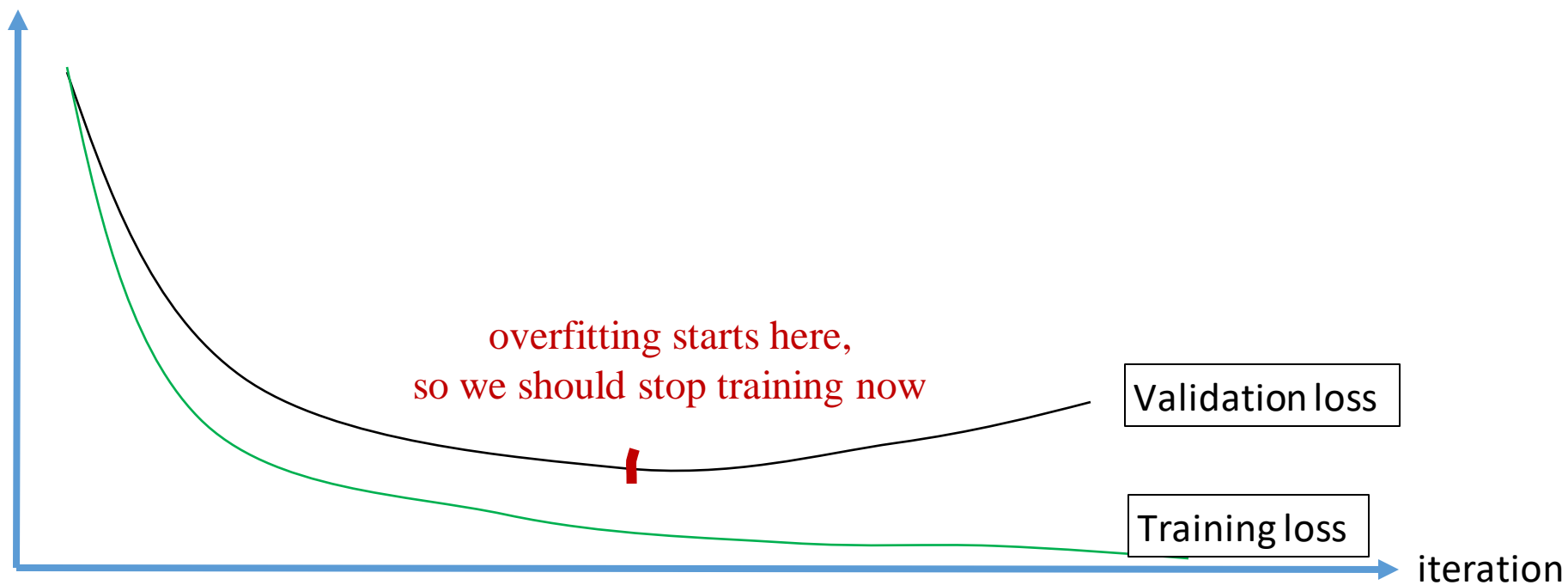


Regularization

** we will discuss this in much more details in subsequent lectures as well as additional regularization techniques

- $J_{reg} = J + \frac{\lambda}{2} ||\Phi||^2$, where Φ is the flatten vector of all parameters
- By minimizing the objective,
 - The parameters will be regularized/constrained to be small values
 - reduce overfitting
- For any $w \in \Phi$
 - $\frac{\partial J_{reg}}{\partial w} = \frac{\partial J}{\partial w} + \lambda w$

Regularization --- early stopping



You can save check points every a few iterations.
Then you can use the best check point before stopping.
Every check point can be deployed.

Summary

- Backpropagation algorithm
 - A modular way to compute the gradient of the loss w.r.t parameters
 - Based on chain rules and matrix calculus
- Advanced gradient descent algorithms
 - SGD, mini-batch SGD
 - Momentum, RMSProp, Adam
- Some training tricks
 - Parameter initialization
 - Data normalization
 - Regularization and early stopping