

Introduction to Deep Learning

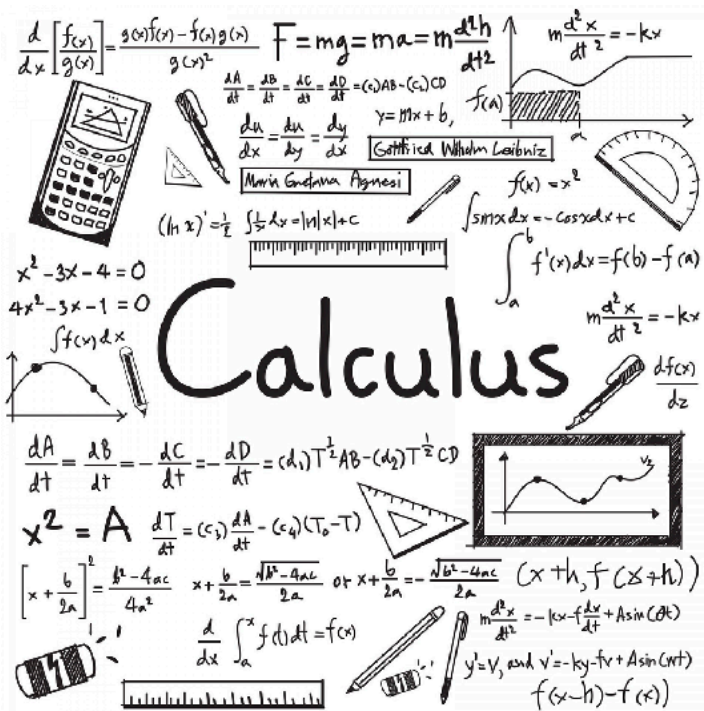
3. Gradient and Auto Differentiation

STAT 157, Spring 2019, UC Berkeley

Alex Smola and Mu Li

courses.d2l.ai/berkeley-stat-157

Matrix



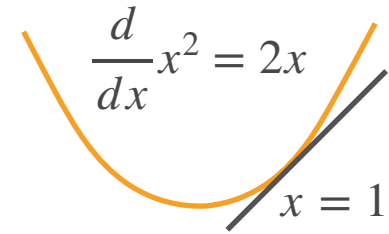
Review Scalar Derivative

y	a	x^n	$\exp(x)$	$\log(x)$	$\sin(x)$
$\frac{dy}{dx}$	0	nx^{n-1}	$\exp(x)$	$\frac{1}{x}$	$\cos(x)$

a is not a function of x

y	$u + v$	uv	$y = f(u), u = g(x)$
$\frac{dy}{dx}$	$\frac{du}{dx} + \frac{dv}{dx}$	$\frac{du}{dx}v + \frac{dv}{dx}u$	$\frac{dy}{du} \frac{du}{dx}$

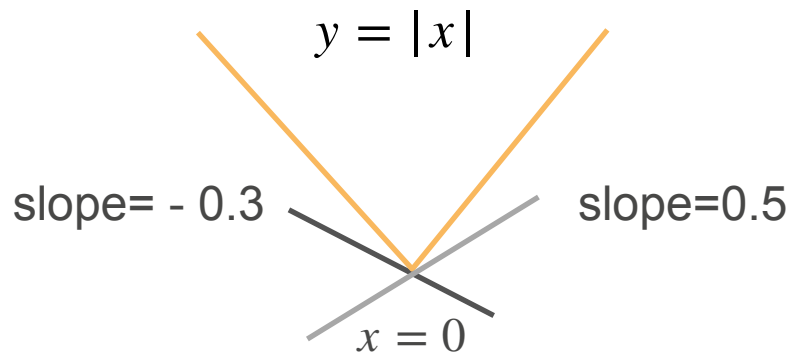
Derivative is the slope of the tangent line



The slope of the tangent line is 2

Subderivative

- Extend derivative to non-differentiable cases



Another example:

$$\frac{\partial}{\partial x} \max(x, 0) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \\ a & \text{if } x = 0, \quad a \in [0, 1] \end{cases}$$

$$\frac{\partial |x|}{\partial x} = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ a & \text{if } x = 0, \quad a \in [-1, 1] \end{cases}$$

Gradients

- Generalize derivatives into vectors

		Scalar	Vector
		x	\mathbf{x}
Scalar	y	$\frac{\partial y}{\partial x}$	$\frac{\partial y}{\partial \mathbf{x}}$
Vector	\mathbf{y}	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$

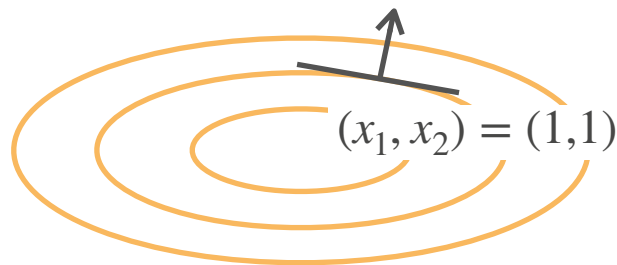
$$\partial y / \partial \mathbf{x}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \frac{\partial y}{\partial \mathbf{x}} = \left[\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \dots, \frac{\partial y}{\partial x_n} \right]$$

	x	\mathbf{x}
y	$\frac{\partial y}{\partial x}$	$\frac{\partial y}{\partial \mathbf{x}}$
\mathbf{y}	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$

$$\frac{\partial}{\partial \mathbf{x}} x_1^2 + 2x_2^2 = [2x_1, 4x_2]$$

Direction (2, 4), perpendicular to the contour lines



Examples

y	a	au	$\text{sum}(\mathbf{x})$	$\ \mathbf{x}\ ^2$
$\frac{\partial y}{\partial \mathbf{x}}$	$\mathbf{0}^T$	$a \frac{\partial u}{\partial \mathbf{x}}$	$\mathbf{1}^T$	$2\mathbf{x}^T$

a is not a function of \mathbf{x}

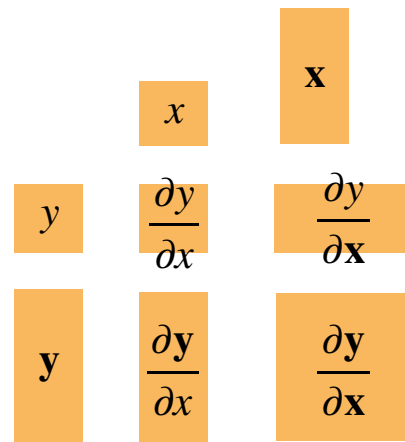
$\mathbf{0}$ and $\mathbf{1}$ are vectors

y	$u + v$	uv	$\langle \mathbf{u}, \mathbf{v} \rangle$
$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial u}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}}$	$\frac{\partial u}{\partial \mathbf{x}} v + \frac{\partial v}{\partial \mathbf{x}} u$	$\mathbf{u}^T \frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^T \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$

$\partial \mathbf{y} / \partial x$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$



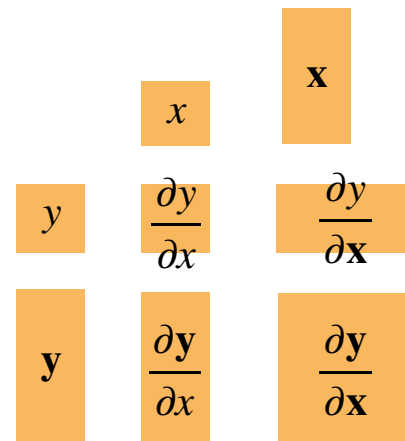
$\partial y / \partial \mathbf{x}$ is a row vector, while $\partial \mathbf{y} / \partial x$ is a column vector

It is called numerator-layout notation. The reversed version is called denominator-layout notation

$\partial \mathbf{y} / \partial \mathbf{x}$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \mathbf{x}} \\ \frac{\partial y_2}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial y_m}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1}, \frac{\partial y_1}{\partial x_2}, \dots, \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1}, \frac{\partial y_2}{\partial x_2}, \dots, \frac{\partial y_2}{\partial x_n} \\ \vdots \\ \frac{\partial y_m}{\partial x_1}, \frac{\partial y_m}{\partial x_2}, \dots, \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$



Examples

y	\mathbf{a}	\mathbf{x}	\mathbf{Ax}	$\mathbf{x}^T \mathbf{A}$
$\frac{\partial y}{\partial \mathbf{x}}$	$\mathbf{0}$	\mathbf{I}	\mathbf{A}	\mathbf{A}^T

$$\mathbf{x} \in \mathbb{R}^n, \quad \mathbf{y} \in \mathbb{R}^m, \quad \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{m \times n}$$

a , \mathbf{a} and \mathbf{A} are not functions of \mathbf{x}

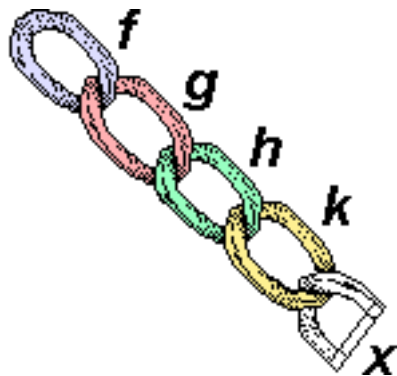
$\mathbf{0}$ and \mathbf{I} are matrices

y	$a\mathbf{u}$	\mathbf{Au}	$\mathbf{u} + \mathbf{v}$
$\frac{\partial y}{\partial \mathbf{x}}$	$a \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\mathbf{A} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$

Generalize to Matrices

		Scalar	Vector	Matrix
		x (1,)	\mathbf{x} (n,1)	\mathbf{X} (n,k)
Scalar	y (1,)	$\frac{\partial y}{\partial x}$ (1,)	$\frac{\partial y}{\partial \mathbf{x}}$ (1,n)	$\frac{\partial y}{\partial \mathbf{X}}$ (k,n)
Vector	\mathbf{y} (m,1)	$\frac{\partial \mathbf{y}}{\partial x}$ (m,1)	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ (m,n)	$\frac{\partial \mathbf{y}}{\partial \mathbf{X}}$ (m,k,n)
Matrix	\mathbf{Y} (m,l)	$\frac{\partial \mathbf{Y}}{\partial x}$ (m,l)	$\frac{\partial \mathbf{Y}}{\partial \mathbf{x}}$ (m,l,n)	$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$ (m,l,k,n)

Chain Rule



Generalize to Vectors

- Chain rule for scalars:

$$y = f(u), u = g(x) \quad \frac{\partial y}{\partial x} = \frac{\partial y}{\partial u} \frac{\partial u}{\partial x}$$

- Generalize to vectors straightforwardly

$$\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial y}{\partial u} \frac{\partial u}{\partial \mathbf{x}}$$

$$(1, n) \quad (1,) \quad (1, n)$$

$$\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial y}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$

$$(1, n) \quad (1, k) \quad (k, n)$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$

$$(m, n) \quad (m, k) \quad (k, n)$$

Example 1

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$

Assume $\mathbf{x}, \mathbf{w} \in \mathbb{R}^n$, $y \in \mathbb{R}$

$$z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$$

Compute $\frac{\partial z}{\partial \mathbf{w}}$

$$\begin{aligned} \frac{\partial z}{\partial \mathbf{w}} &= \frac{\partial z}{\partial b} \frac{\partial b}{\partial a} \frac{\partial a}{\partial \mathbf{w}} \\ &= \frac{\partial b^2}{\partial b} \frac{\partial a - y}{\partial a} \frac{\partial \langle \mathbf{x}, \mathbf{w} \rangle}{\partial \mathbf{w}} \\ &= 2b \cdot 1 \cdot \mathbf{x}^T \\ &= 2 (\langle \mathbf{x}, \mathbf{w} \rangle - y) \mathbf{x}^T \end{aligned}$$

Decompose

$$a = \langle \mathbf{x}, \mathbf{w} \rangle$$

$$b = a - y$$

$$z = b^2$$

Example 2

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$

Assume $\mathbf{X} \in \mathbb{R}^{m \times n}$, $\mathbf{w} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$

$$z = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

Compute $\frac{\partial z}{\partial \mathbf{w}}$

$$\begin{aligned} \frac{\partial z}{\partial \mathbf{w}} &= \frac{\partial z}{\partial \mathbf{b}} \frac{\partial \mathbf{b}}{\partial \mathbf{a}} \frac{\partial \mathbf{a}}{\partial \mathbf{w}} \\ &= \frac{\partial \|\mathbf{b}\|^2}{\partial \mathbf{b}} \frac{\partial \mathbf{a} - \mathbf{y}}{\partial \mathbf{a}} \frac{\partial \mathbf{X}\mathbf{w}}{\partial \mathbf{w}} \\ &= 2\mathbf{b}^T \times \mathbf{I} \times \mathbf{X} \\ &= 2(\mathbf{X}\mathbf{w} - \mathbf{y})^T \mathbf{X}^T \end{aligned}$$

Decompose

$$\begin{aligned} \mathbf{a} &= \mathbf{X}\mathbf{w} \\ \mathbf{b} &= \mathbf{a} - \mathbf{y} \\ z &= \|\mathbf{b}\|^2 \end{aligned}$$

Auto Differentiation



Auto Differentiation (AD)

- AD evaluates gradients of a function specified by a program at given values
- AD differs to
 - Symbolic differentiation

```
In[1]:= D[4 x^3 + x^2 + 3, x]
```

```
Out[1]= 2 x + 12 x^2
```

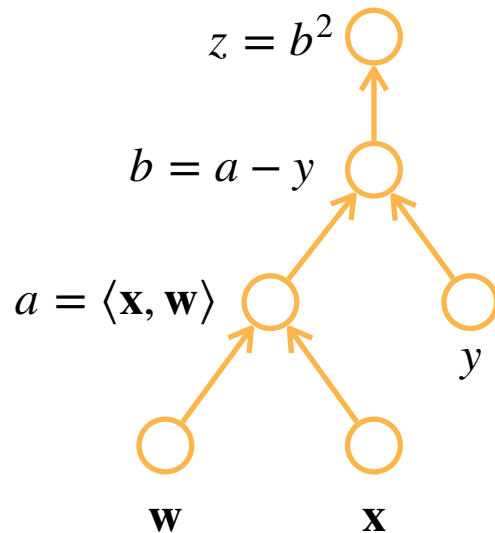
- Numerical differentiation

$$\frac{\partial f(x)}{\partial x} = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

Computation Graph

- Decompose into primitive operations
- Build a directed acyclic graph to present the computation

Assume $z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$



Computation Graph

- Decompose into primitive operations
- Build a directed acyclic graph to present the computation
- Build explicitly
 - Tensorflow/Theano/MXNet

```
from mxnet import sym
```

```
a = sym.var()
```

```
b = sym.var()
```

```
c = 2 * a + b
```

```
# bind data into a and b later
```

Computation Graph

- Decompose into primitive operations
- Build a directed acyclic graph to present the computation
- Build explicitly
 - Tensorflow/Theano/MXNet
- Build implicitly though tracing
 - PyTorch/MXNet

```
from mxnet import autograd, nd
```

```
with autograd.record():  
    a = nd.ones((2,1))  
    b = nd.ones((2,1))  
    c = 2 * a + b
```

Two Modes

- By chain rule
$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial u_n} \frac{\partial u_n}{\partial u_{n-1}} \dots \frac{\partial u_2}{\partial u_1} \frac{\partial u_1}{\partial x}$$

- Forward accumulation

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial u_n} \left(\frac{\partial u_n}{\partial u_{n-1}} \left(\dots \left(\frac{\partial u_2}{\partial u_1} \frac{\partial u_1}{\partial x} \right) \right) \right)$$

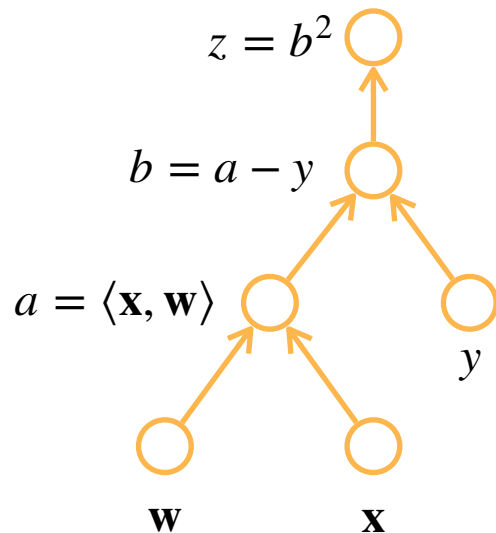
- Reverse accumulation (a.k.a Backpropagation)

$$\frac{\partial y}{\partial x} = \left(\left(\left(\frac{\partial y}{\partial u_n} \frac{\partial u_n}{\partial u_{n-1}} \right) \dots \right) \frac{\partial u_2}{\partial u_1} \right) \frac{\partial u_1}{\partial x}$$

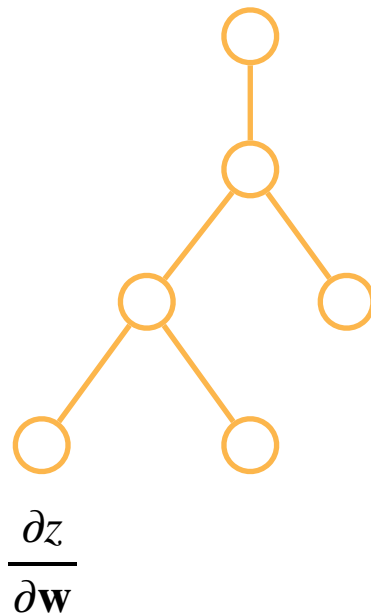
Reverse Accumulation

Assume $z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$

Forward

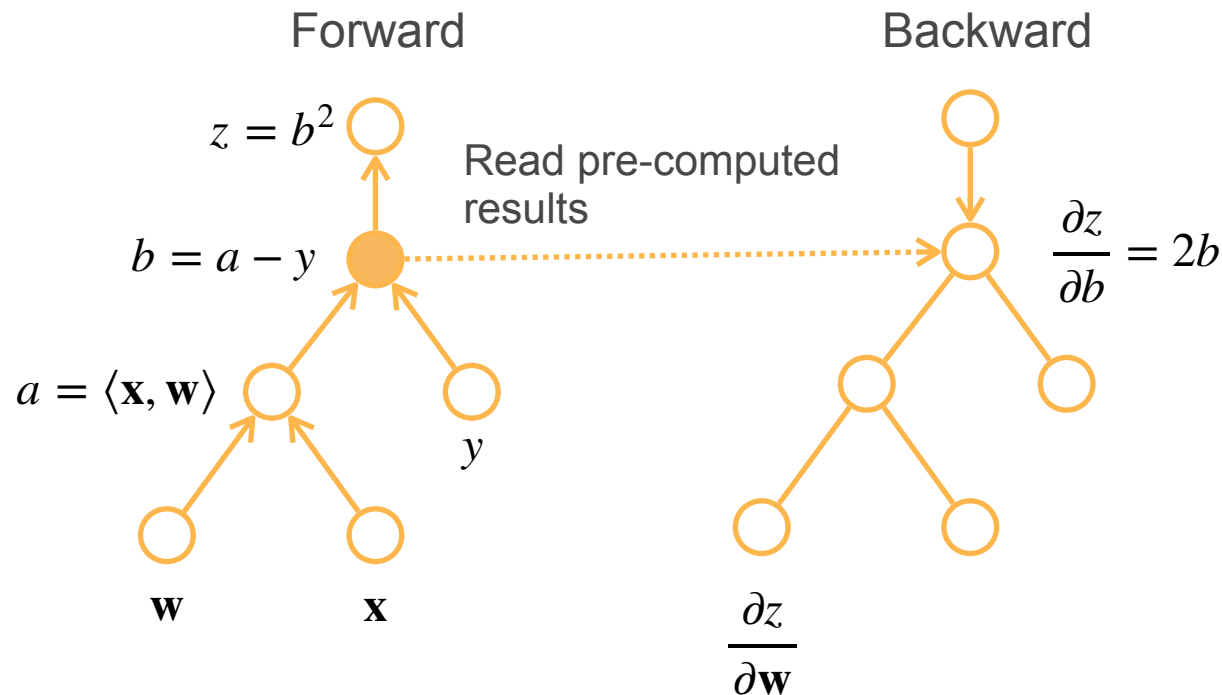


Backward



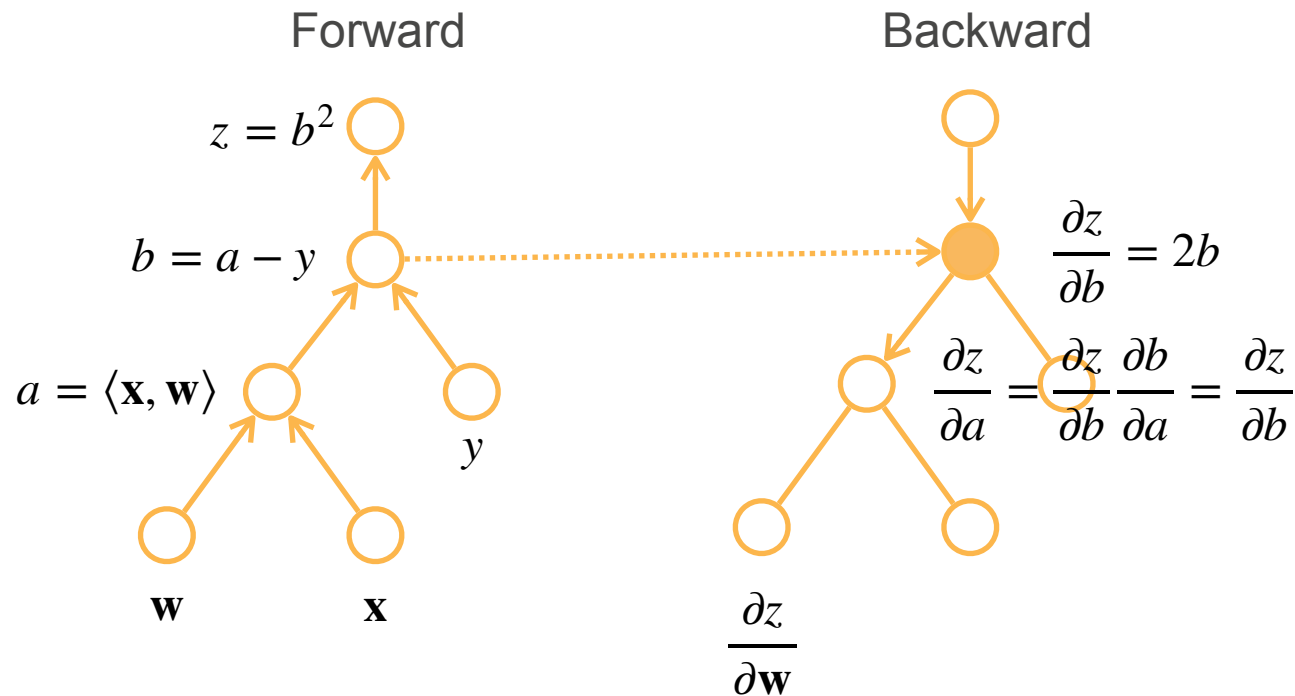
Reverse Accumulation

Assume $z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$



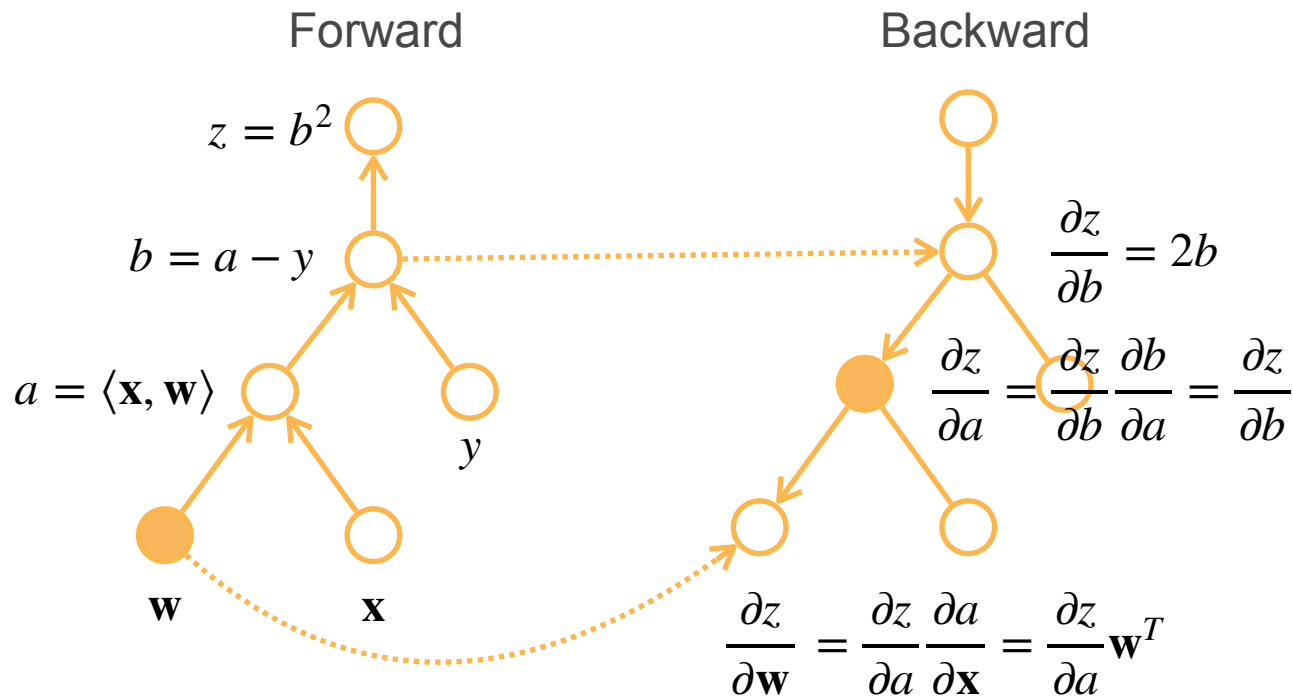
Reverse Accumulation

Assume $z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$



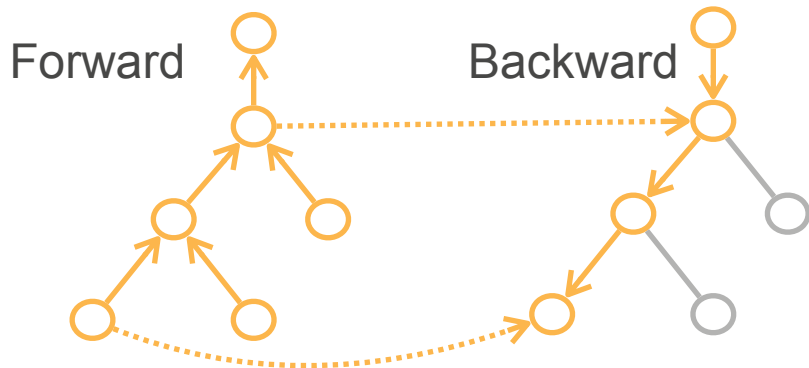
Reverse Accumulation

Assume $z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$



Reverse Accumulation Summary

- Build a computation graph
- Forward: Evaluate the graph, store intermediate results
- Backward: Evaluate the graph in a reversed order
 - Eliminate paths not needed



Complexities

- Computational complexity: $O(n)$, n is #operations, to compute all derivatives
 - Often similar to the forward cost
- Memory complexity: $O(n)$, needs to record all intermediate results in the forward pass
- Compare to forward accumulation:
 - $O(n)$ time complexity to compute one gradient, $O(n*k)$ to compute gradients for k variables
 - $O(1)$ memory complexity

[Advanced] Rematerialization

- Memory is bottleneck for backward accumulation
 - Linear to #layers and batch size
 - Limited GPU memory (32GB max)
- Trade computation for memory
 - Save a part of intermediate results
 - Recompute the rest when needed

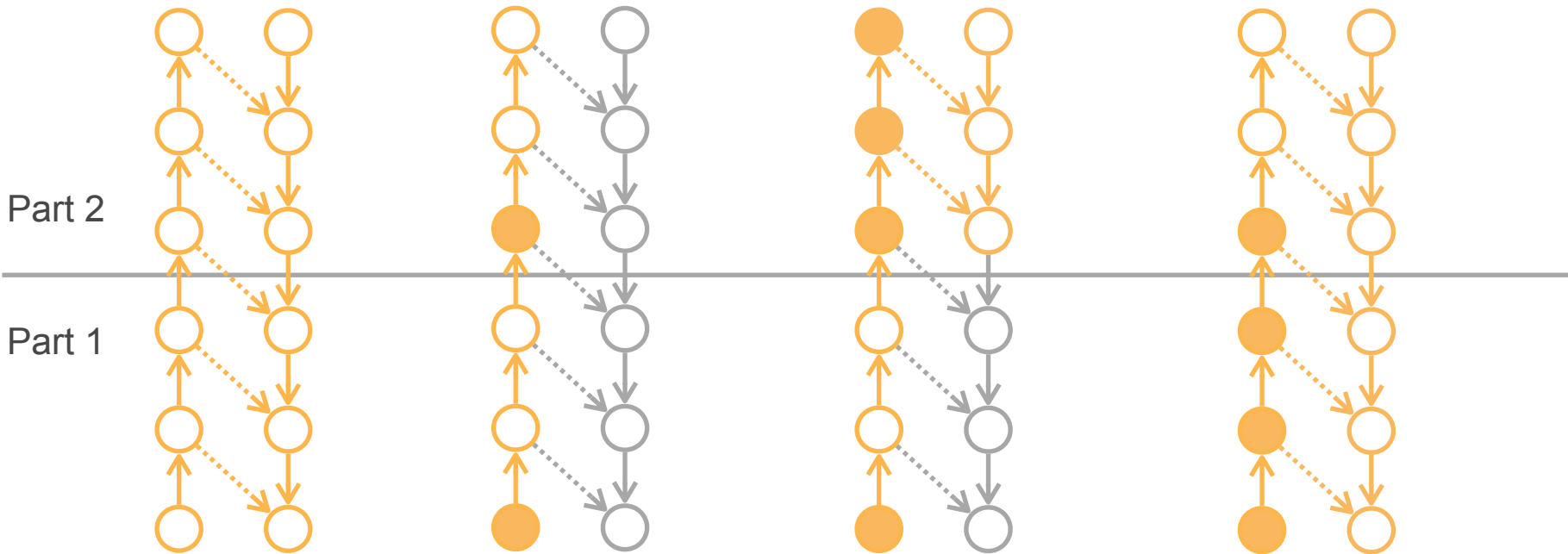
Rematerialization

Forward Backward

Only store the head
result in each part

Recompute the
rest in part 2

Recompute the
rest in part 1



Complexities

- An additional forward pass
- Assume m parts, then $O(m)$ for head results, $O(n/m)$ to store one part's results
 - Choose $m = \sqrt{n}$ then the memory complexity is $O(\sqrt{n})$
- Applying to deep neural networks
 - Only throw away simple layers, e.g. activation, often <30% additional overhead
 - Train 10x larger networks, or 10x large batch size

Autograd in MXNet

https://d2l.ai/chapter_crashcourse/autograd.html

Limitations

- Does not support every operations
 - Indexing
 - Inplace
- Not smart enough to get numerical stable results
 - Homework