3 导数、反向传播和复杂度

概要

- 矩阵微积分
- 链式法则
- 自动微分法
- 反向传播

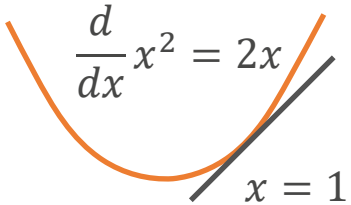
矩阵微积分

标量求导回顾

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

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

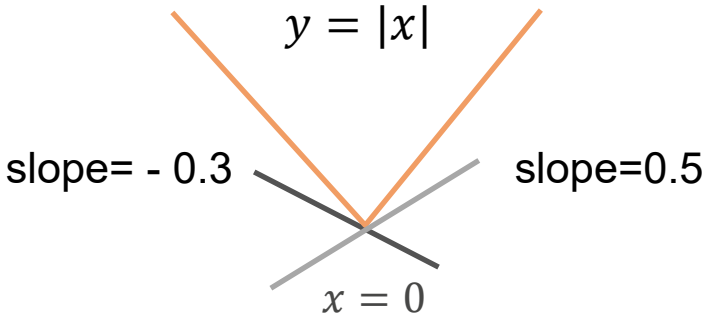
导数是切线的斜率



切线的斜率为2

次导数

- 不可求导情况下的导数
  - Example 1:



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

- Example 2:

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

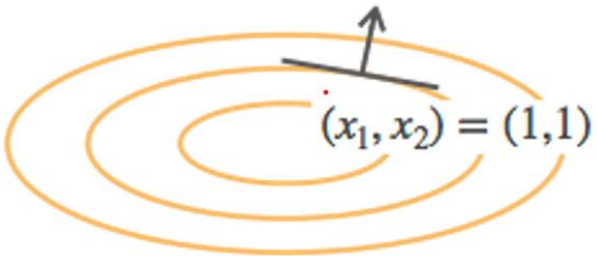
梯度

- 矢量求导推广

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

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

- $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \frac{\partial y}{\partial \mathbf{x}} = \left[\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \ldots, \frac{\partial y}{\partial x_n}\right]$
- $\frac{\partial}{\partial \mathbf{x}}(x_1^2 + 2x_2^2) = [2x_1, 4x_2]$



例子

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

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

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

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

- $\partial \mathbf{y}/\partial x$ 分子布局 (numerator-layout 或 Jacobian formulation)， 是行矢量
- $\partial y/\partial \mathbf{x}$ 分母布局（denominator-layout 或 Hessian formulation），是列矢量

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

- $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$ $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{m \times n}$

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

例子

- $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m, \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{m \times n}$,
- $a, \mathbf{a}$ 和 $\mathbf{A}$ 不是关于$\mathbf{x}$的函数
- $\mathbf{0}$ 和 $\mathbf{I}$ 为矩阵

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

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

推广到矩阵

		标量	矢量	矩阵
		$x\,(1,)$	$\mathbf{x}\,(n,1)$	$\mathbf{X}(n,k)$
标量	$y\,(1,)$	$\frac{\partial y}{\partial x}\,(1,)$	$\frac{\partial y}{\partial \mathbf{x}}\,(1,n)$	$\frac{\partial y}{\partial \mathbf{X}}\,(k,n)$
矢量	$\mathbf{y}\,(m,1)$	$\frac{\partial \mathbf{y}}{\partial x}\,(m,1)$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\,(m,n)$	$\frac{\partial \mathbf{y}}{\partial \mathbf{X}}\,(m,k,n)$
矩阵	$\mathbf{Y}\,(m,l)$	$\frac{\partial \mathbf{Y}}{\partial x}\,(m,l)$	$\frac{\partial \mathbf{Y}}{\partial \mathbf{x}}\,(m,l,n)$	$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}\,(m,l,k,n)$

链式法则

链式法则

- 链式法则 – 标量：

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

- 链式法则 – 矢量：

$$\underset{(1,n)}{\frac{\partial y}{\partial \mathbf{x}}} = \underset{(1,)}{\frac{\partial y}{\partial u}}\underset{(1,n)}{\frac{\partial u}{\partial \mathbf{x}}} \qquad \underset{(1,n)}{\frac{\partial y}{\partial \mathbf{x}}} = \underset{(1,k)}{\frac{\partial y}{\partial \mathbf{u}}}\underset{(k,n)}{\frac{\partial \mathbf{u}}{\partial \mathbf{x}}} \qquad \underset{(m,n)}{\frac{\partial \mathbf{y}}{\partial \mathbf{x}}} = \underset{(m,k)}{\frac{\partial \mathbf{y}}{\partial \mathbf{u}}}\underset{(k,n)}{\frac{\partial \mathbf{u}}{\partial \mathbf{x}}}$$

例1

- 假设　$\mathbf{x}, \mathbf{w} \in \mathbb{R}^n$, $y \in \mathbb{R}$ $z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$

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

例1

- 假设 $\mathbf{x}, \mathbf{w} \in \mathbb{R}^n$, $y \in \mathbb{R}$ $z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$
  - 计算 $\frac{\partial z}{\partial \mathbf{w}}$
  - 分解
    - $a = \langle \mathbf{x}, \mathbf{W} \rangle$
    - $b = a - y$
    - $z = b^2$
- 求偏导
- $\frac{\partial z}{\partial \mathbf{w}} = \frac{\partial z}{\partial b}\frac{\partial b}{\partial a}\frac{\partial a}{\partial \mathbf{w}} = \frac{\partial b^2}{\partial b}\frac{\partial a - y}{\partial a}\frac{\partial \langle \mathbf{x}, \mathbf{w} \rangle}{\partial \mathbf{w}} = 2b \cdot 1 \cdot \mathbf{x}^T = 2(\langle \mathbf{x}, \mathbf{w} \rangle - y)\mathbf{x}^T$

例2

假设 $\mathbf{X} \in \mathbb{R}^{m \times n}, \quad \mathbf{w} \in \mathbb{R}^{n}, \quad \mathbf{y} \in \mathbb{R}^{m} \qquad z = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

计算 $\dfrac{\partial z}{\partial \mathbf{w}}$

分解

$$\mathbf{a} = \mathbf{X}\mathbf{w}$$
$$\mathbf{b} = \mathbf{a} - \mathbf{y}$$
$$z = \|\mathbf{b}\|^2$$
$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$

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

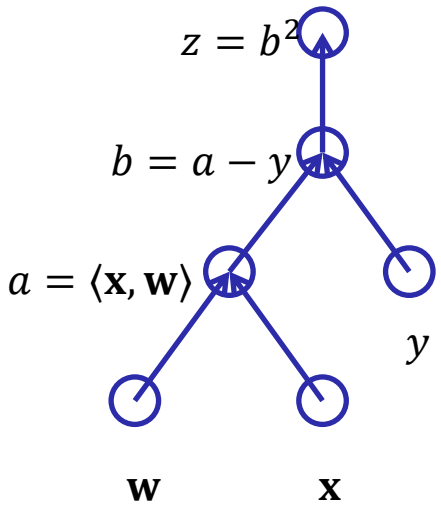
自动微分法

自动微分（AD）

- 自动微分(AD)将符号微分法应用于最基本的算子，然后代入数值，应用于整个函数
- 其它常见微分法
  - 符号微分法
    - $\mathrm{In}[1] := \mathrm{D}[4\mathrm{x}^3 + \mathrm{x}^2 + 3, \mathrm{x}]$
    - $\mathrm{Out}[1] = 2\mathrm{x} + 12\mathrm{x}^2$
  - 数值微分法
    - $\frac{\partial f(x)}{\partial x} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$

计算图

- 分解成最基本的方程
- 构造有向无环图来表示运算
- 假设$z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$



计算图

- 分解成最基本的方程
- 构造有向无环图来表示运算
- 显性构造Tensorflow/Theano/MXNet
- 隐性构造
- PyTorch/MXNet

```
from mxnet import autograd, nd

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

两种微分模式

- 正向传播：一个输入的变化，如何影响所有输出
  - 从输入（某叶节点）开始，前向传播计算节点对输入的微分
  - 首先前向传播计算图中每个节点的值
  - 与函数求值顺序相同，求导所需的函数值可以和对应的微分一起求出来
  - 【缺点】每次前向计算只能计算对一个自变量的偏导。n 个输入的梯度需要 n 遍计算

链式法则：$\frac{\partial y}{\partial x} = \frac{\partial u_n}{\partial u_{n-1}} \cdots \frac{\partial u_2}{\partial u_1} \frac{\partial u_1}{\partial x}$

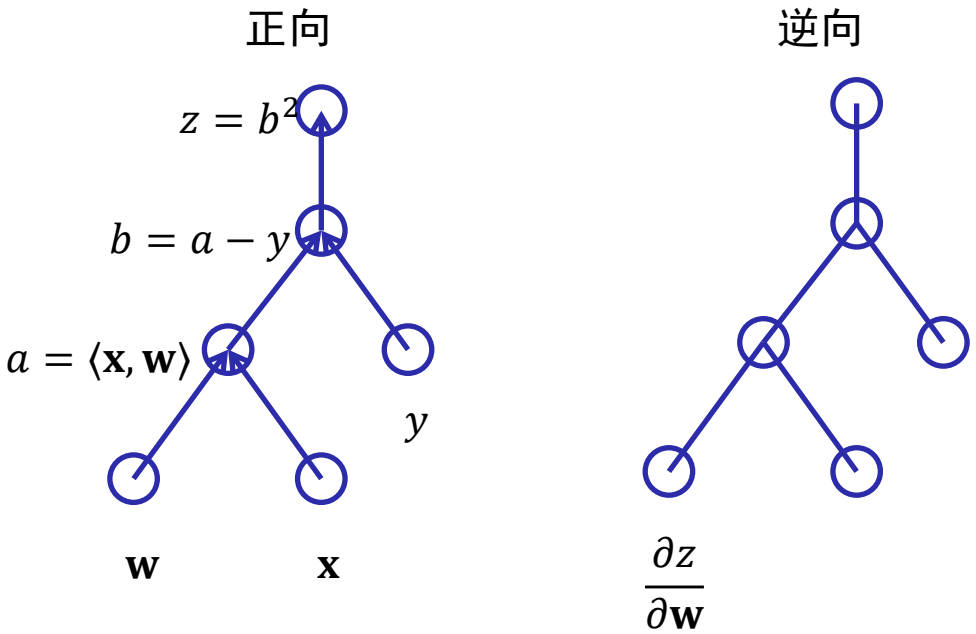
$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial u_n}\left(\frac{\partial u_n}{\partial u_{n-1}}\left(\cdots\left(\frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial x}\right)\right)\right)$$

- 反向传播：一个输出的变化，如何由每个输入贡献
  - 从输出（根节点）开始，反向传播计算输出对每个节点的偏导
  - 一次反向传输计算出所有偏导数，中间的偏导数计算只需计算一次，减少了重复计算的工作量
  - 【缺点】需要额外的数据结构记录正向过程的计算操作，用于反向使用

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

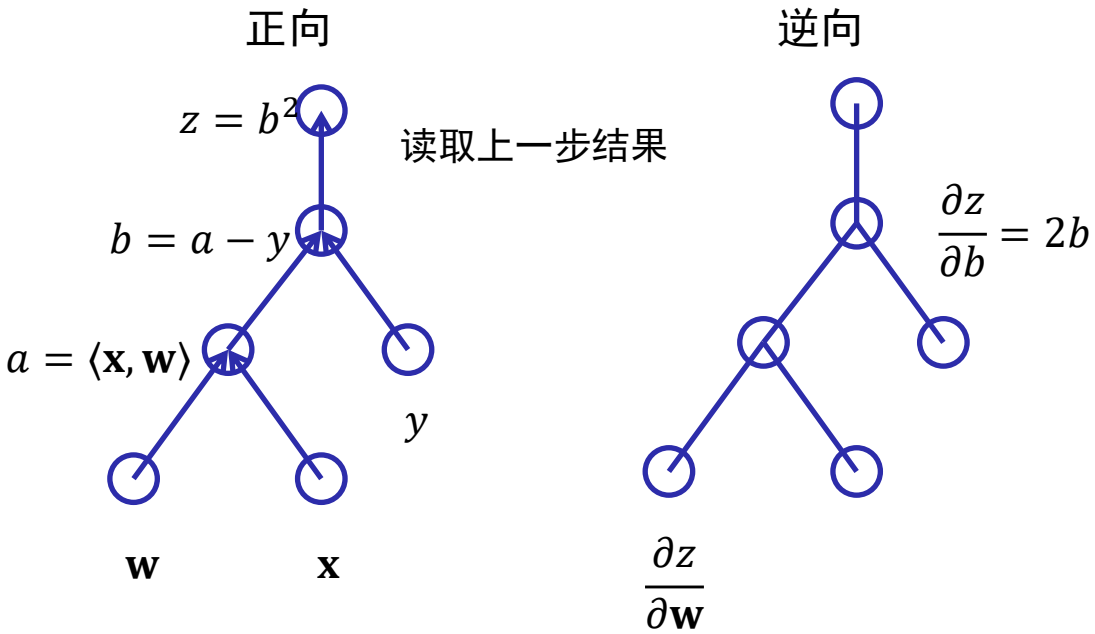
反向传播

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



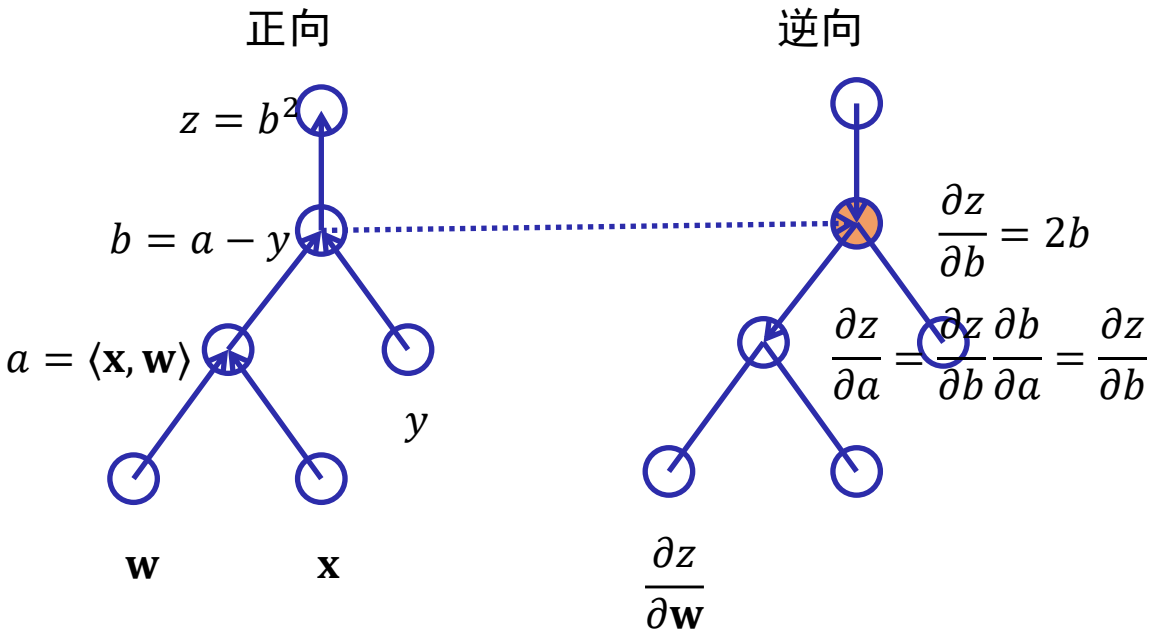
反向传播

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



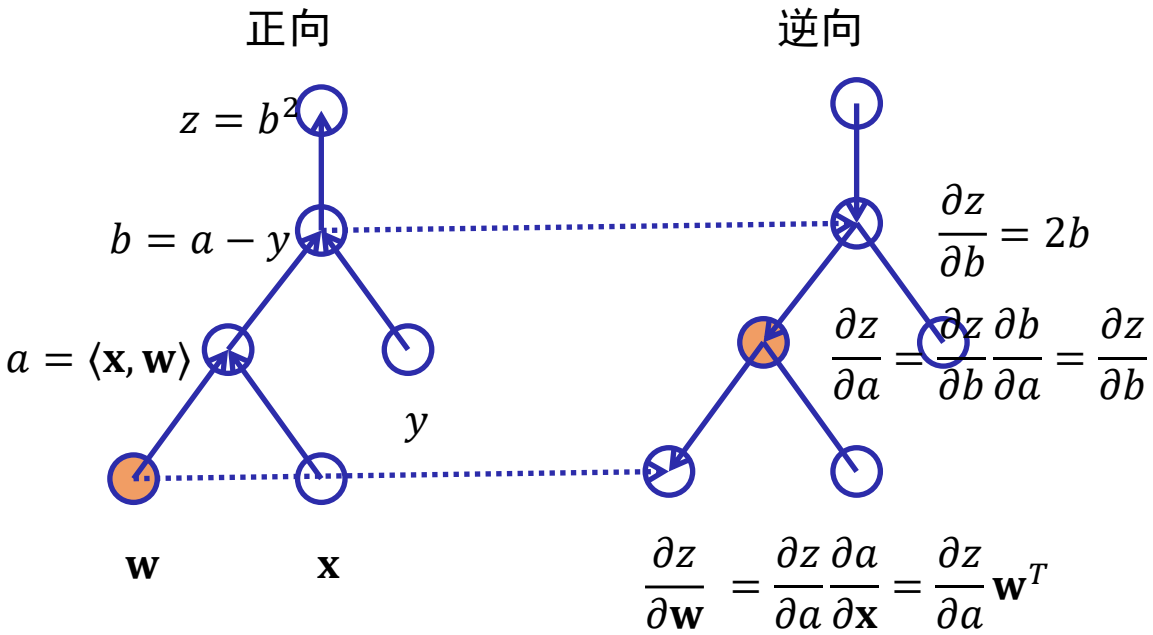
反向传播

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



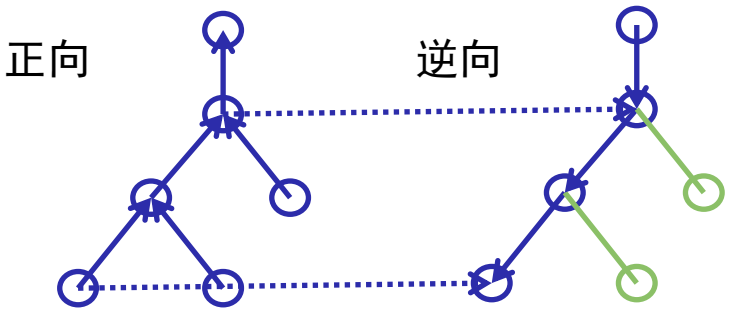
反向传播

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



反向传播总结

- 创建一个计算图
- 正向：计算有向无环图，储存中间值
- 反向：逆向计算有向无环图
  - 减少不需要的图



复杂度

- O(n)，n 为计算次数

- 反向传播复杂度：
  - 时间复杂度：O(n)，计算所有导数，基本上与正向复杂度一致
  - 内存复杂度：O(n)，需要储存所有正向计算的中间值
- 对比正向传播:
  - 时间复杂度：O(n)，计算 k 个变量的导数为 O(n*k)
  - 内存复杂度：O(1)

[拓展] 再具体化（re-materialization）

- 内存是逆向传播的瓶颈
  - 随着层数和批量大小线性增长
  - 有限 GPU 内存（最多32GB）
- 用算力换内存
  - 只保存一部分中间计算值
  - 当需要时重新计算未保存中间值

总结

- 矩阵微积分
- 链式法则
- 自动微分法
- 反向传播