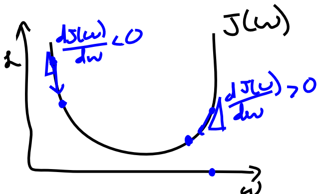
## Gradient Descent

$\hat{y} = \sigma(w^T x + b)$, $\sigma(z) = \frac{1}{1+e^{-z}}$

$J(w,b) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m}\sum_{i=1}^{m} y\log\hat{y} + (1-y)\log(1-\hat{y})$

## Want to find w,b that minimize J(w,b)




learning rate 'alpha'

Repeat {
$w := w - \alpha \frac{dJ(w)}{dw}$
}

change in J(w) over change in w

'dw'

$w := w - \alpha dw$ ← derivative

J(w,b) $w := w - \alpha \frac{dJ(w,b)}{dw}$

$b := b - \alpha \frac{dJ(w,b)}{db}$

definition of derivative is the slope of the curve

$\partial$ ← partial derivative (used for more than one variable)

$d$ ← one variable

Technically should be $\frac{\partial J(w,b)}{\partial w}$ but doesn't actually matter

V10

## Derivatives

$f(x) = cx^n$

$\frac{d}{dx} f(x) = cnx^{n-1}$

eg. $f(x) = 3x$, $\frac{d}{dx} f(x) = 3$ | $f(x) = x^3$, $\frac{d}{dx} f(x) = 3x^2$ | $f(x) = 3x^3$, $\frac{d}{dx} f(x) = 9x^2$

$f(x) = \log_e(x) = \ln(x)$

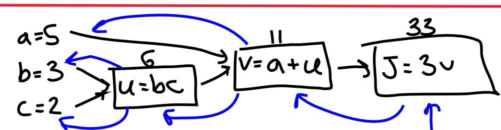
$\frac{d}{dx} f(x) = \frac{1}{x}$

V11 + V12

## Computation Graph

$J(a,b,c) = 3(\underbrace{a + \underbrace{bc}_{u}}_{v})$

$u = bc$
$v = a + u$
$J = 3v$

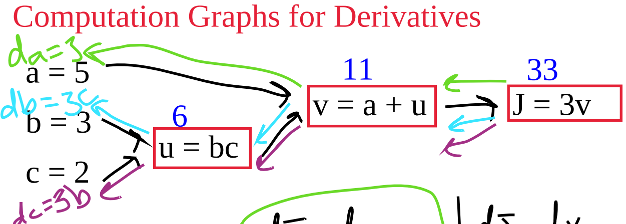



Computation graph going forward

Derivatives go back

V13

## Computation Graphs for Derivatives




$\frac{dJ}{dv} = 3$ → $\frac{dJ}{dv}\frac{dv}{da} = 3$ ($3$, $1$) | $\frac{dJ}{dv}\frac{dv}{du} = 3$ ($3$, $1$)

$\frac{dJ}{dv}\frac{dv}{du}\frac{du}{db} = 3c$ ($3$, $1$, $c$) | $\frac{dJ}{dv}\frac{dv}{du}\frac{du}{dc} = 3b$ ($3$, $1$, $b$)

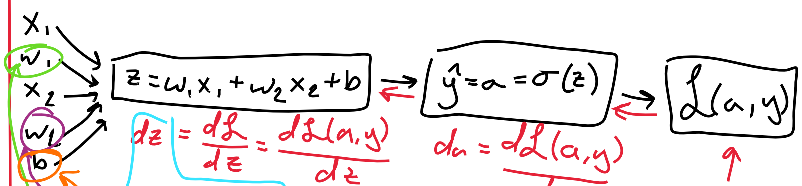
V14

## Logistic Regression Gradient Descent

$z = w^T x + b$

$\hat{y} = a = \sigma(z)$

$\mathcal{L}(a,y) = -(y\log(a) + (1-y)\log(1-a))$




$dz = \frac{d\mathcal{L}}{dz} = \frac{d\mathcal{L}(a,y)}{dz}$

$= a - y$

$= \boxed{\frac{d\mathcal{L}}{da}} \cdot \frac{da}{dz}$

$da = \frac{d\mathcal{L}(a,y)}{da}$

$= -\frac{y}{a} + \frac{1-y}{1-a}$

$\frac{da}{dz} = a(1-a)$

$\frac{\partial \mathcal{L}}{\partial w_1} = \text{"}dw_1\text{"} = x_1 dz$ $\quad dw_2 = x_2 \cdot dz$ $\quad db = dz$

### Updates

learning rate alpha

$w_1 := w_1 - \alpha dw_1$
$= w_1 - \alpha(x_1 dz)$
$= w_1 - \alpha(x_1(a-y))$
$= w_1 - \alpha(x_1(\sigma(z)-y))$

$\hat{y}$

$w_2 := w_2 - \alpha dw_2$

$b := b - \alpha db$

V15

## Logistic Regression on m examples

$(x^{(i)}, y^{(i)})$

$dw_1^{(i)}, dw_2^{(i)}, db^{(i)}$

$J(w,b) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(a\ \ ,y)$

$a^{(i)} = \hat{y}^{(i)} = \sigma(z^{(i)}) = \sigma(w^T x^{(i)} + b)$

$\frac{\partial}{\partial w_1} J(w,b) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{\frac{\partial}{\partial w_1}\mathcal{L}(a^{(i)}, y^{(i)})}_{dw_1^{(i)}} \rightarrow (x^{(i)}, y^{(i)})$

Just as the loss function is the average loss, the derivative is the average derivative.

$J=0;\ dw_1=0;\ dw_2=0;\ db=0$

For $i=1$ to $m$

$z^{(i)} = w^T x^{(i)} + b$

$a^{(i)} = \sigma(z^{(i)})$

$J \mathrel{+}= -[y^{(i)}\log a^{(i)} + (1-y^{(i)})\log(1-a^{(i)})]$

$dz^{(i)} = a^{(i)} - y^{(i)}$

$dw_1 \mathrel{+}= x_1^{(i)} dz^{(i)}$

$dw_2 \mathrel{+}= x_2^{(i)} dz^{(i)}$

$n=2$ this is for 2 features otherwise for the whole matrix

$db \mathrel{+}= dz^{(i)}$

$J \mathrel{/}= m$

$dw_1 \mathrel{/}= m;\ dw_2 \mathrel{/}= m;\ db \mathrel{/}= m$

↑ divide equals

$dw_1 = \frac{\partial J}{\partial w}$

$w_1 := w_1 - \alpha dw_1$

$w_2 := w_2 - \alpha dw_2$

$b := b - \alpha db$

V16