Auto differentiation

$f(a) = 4a^2 - 3$

$f'(a) = 8a$

$f'(1.8) = 14.4$

Slope of $f(a)$ at $a = 1.8$

$f(1.8) = 4 \cdot (1.8)^2 - 3 = 9.96$

What should we do to 'a' to reduce $f(a)$?

$f(1.7) = 4 \cdot (1.7)^2 - 3 = 8.56$

$f(a, k) = k \cdot a + 17$

we have a specific value and we want to pick new value for 'a' such that $f(a,k)$ is reduced

$\frac{\partial f(a,k)}{\partial a} = k$

2 - Layer Network

Layer 1:

$$z^{[1]} = A^{[0]} W^{[1]T} + b^{[1]}$$
$$A^{[1]} = \sigma(z^{[1]})$$

↑ sigmoid activation

$$z^{[2]} = A^{[1]} W^{[2]T} + b^{[2]}$$
$$A^{[2]} = \sigma(z^{[2]})$$

↑ we don't have to use the same activation function for each layer

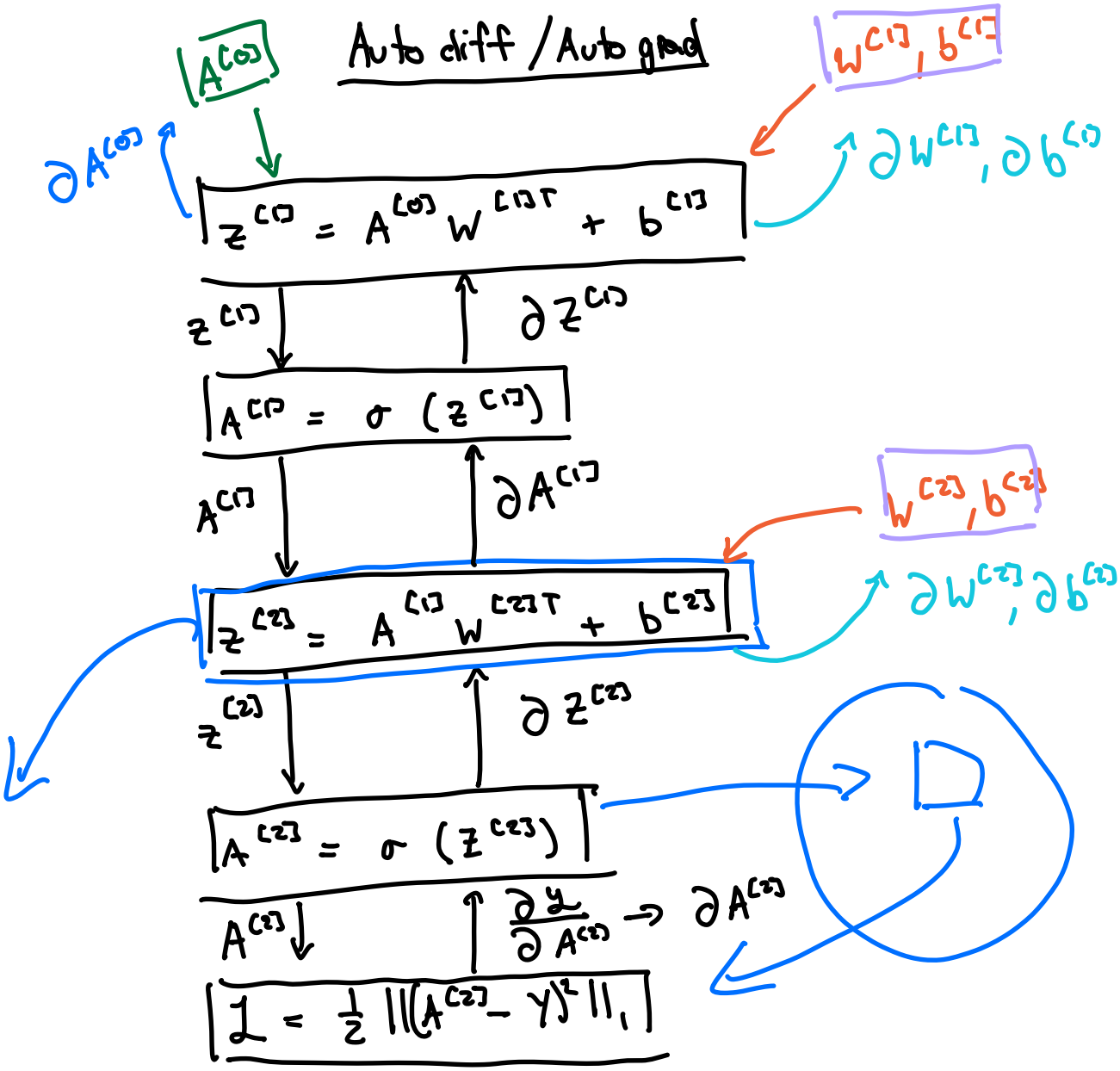
$$A^{[2]} = \sigma(A^{[1]} W^{[2]T} + b^{[2]})$$
$$= \sigma(\sigma(z^{[1]}) W^{[2]T} + b^{[2]})$$
$$= \sigma(\sigma(\sigma(A^{[0]} W^{[1]T} + b^{[1]}) W^{[2]T} + b^{[2]})) W^{[3]T} + b^{[3]}$$

$$\mathcal{L}(\hat{y}, y) = \frac{1}{2}\|(\hat{y} - y)^2\|_1$$

(1) How does $\frac{\partial \mathcal{L}}{\partial W^{[1]}}$ change if we add hidden layers?

(2) How does $\frac{\partial \mathcal{L}}{\partial W^{[1]}}$ change if we change an activation function?

(3) How does $\frac{\partial \mathcal{L}}{\partial W^{[1]}}$ change if we change the loss function?


Auto diff / Auto grad
$A^{[0]}$
$W^{[1]}, b^{[1]}$
$\partial A^{[0]}$
$\partial W^{[1]}, \partial b^{[1]}$
$Z^{[1]} = A^{[0]} W^{[1]T} + b^{[1]}$
$Z^{[1]}$
$\partial Z^{[1]}$
$A^{[1]} = \sigma(Z^{[1]})$
$A^{[1]}$
$\partial A^{[1]}$
$W^{[2]}, b^{[2]}$
$\partial W^{[2]}, \partial b^{[2]}$
$Z^{[2]} = A^{[1]} W^{[2]T} + b^{[2]}$
$Z^{[2]}$
$\partial Z^{[2]}$
$A^{[2]} = \sigma(Z^{[2]})$
$A^{[2]}$
$\frac{\partial L}{\partial A^{[2]}} \rightarrow \partial A^{[2]}$
$L = \frac{1}{2} \| (A^{[2]} - Y)^2 \|_1$