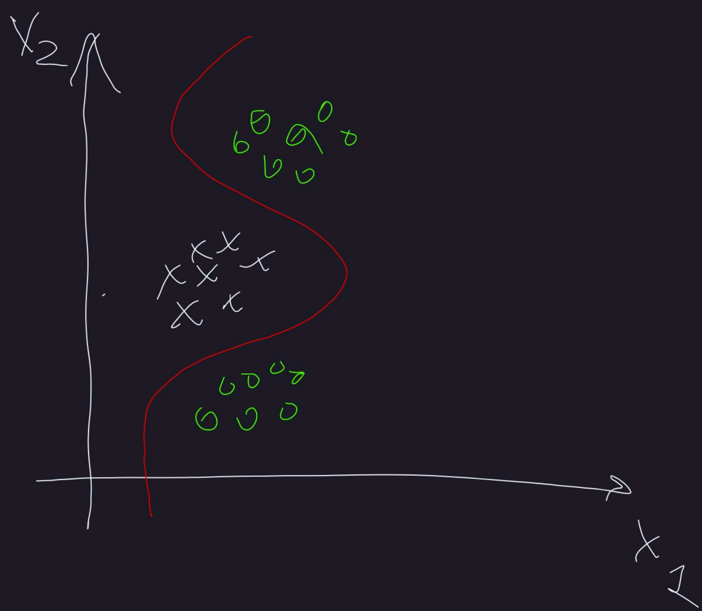
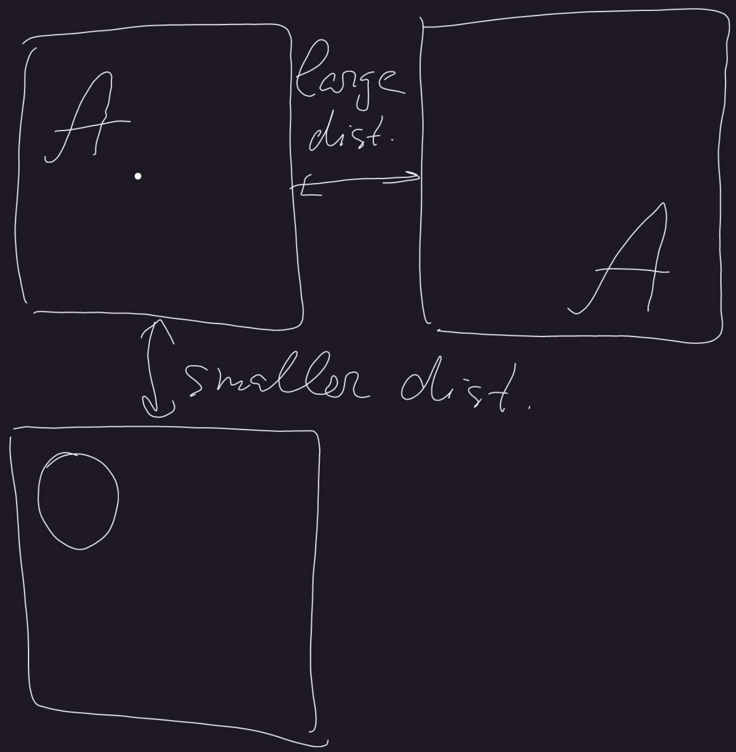
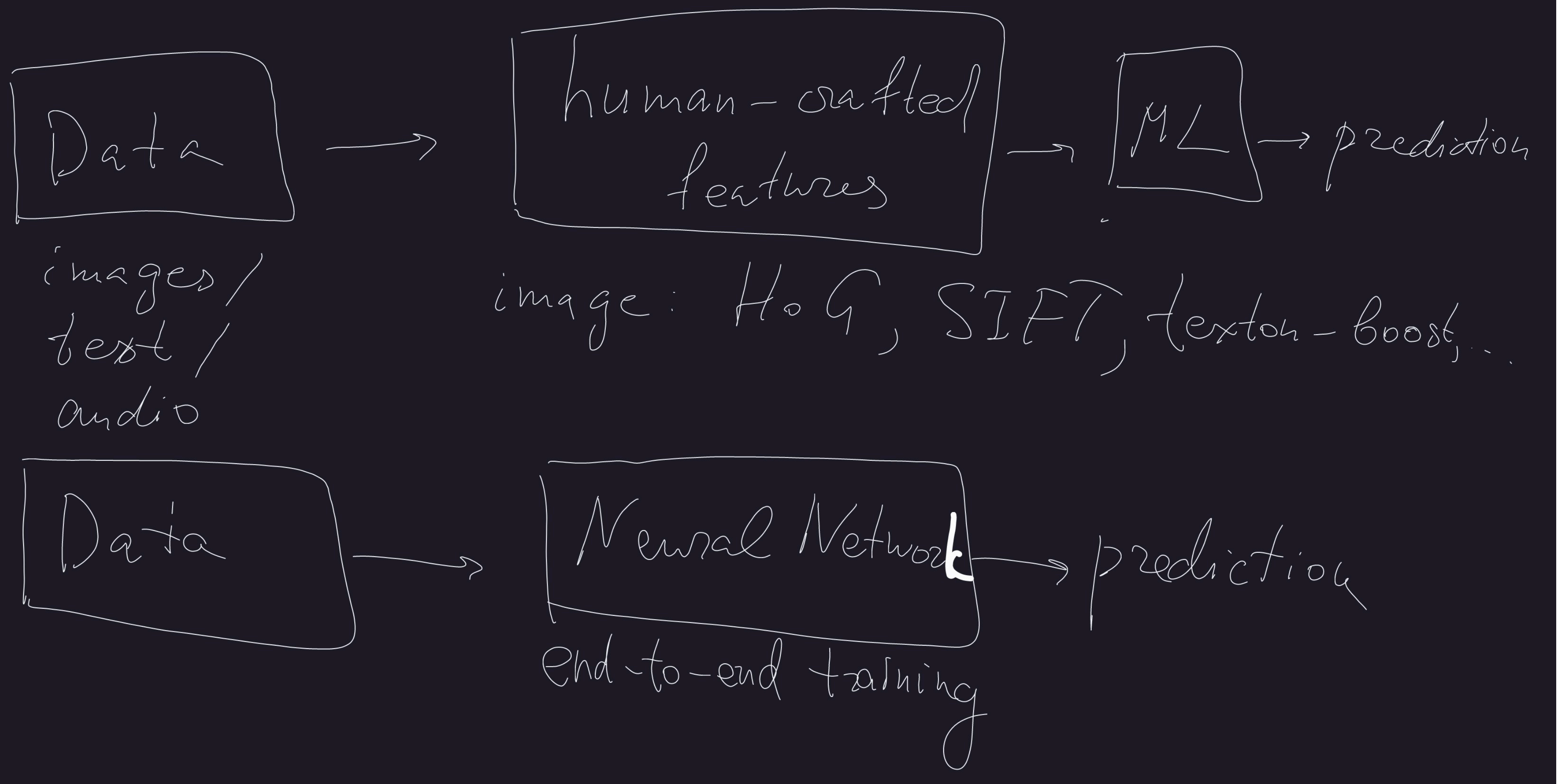

x2
x1


A
large dist.
A
smaller dist.



Data
images / text / audio
human-crafted features
image: HoG, SIFT, texton-boost, ...
ML
prediction
Data
Neural Network
prediction
end-to-end training

# Logistic Regression

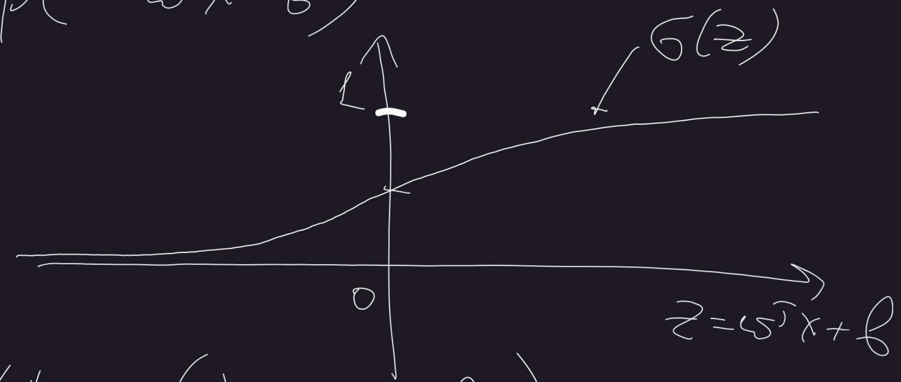
$$\{x_i, y_i\}_{i=1}^{N}, \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, +1\}$$

$$y_i \in \{1, 2, \dots, K\}$$

inference: $$y(x) = \text{sign}(\underset{1\times d}{w^T}\ \underset{d\times 1}{x} + \underset{1\times 1}{b})$$

$$y(x) = \underset{j \in \{1, \dots, K\}}{\arg\max} (w_j^T x + b_j)$$

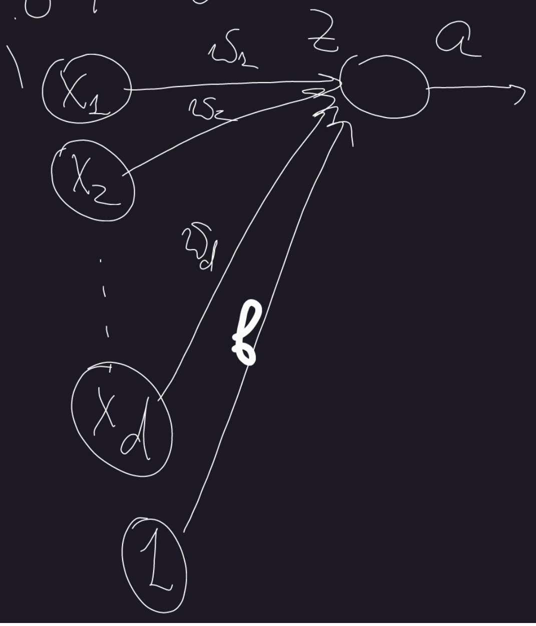
$$p(y=+1 \mid w, x, b) = \frac{1}{1+\exp(-w^T x - b)} = \sigma(w^T x + b)$$




$$p(y=k \mid W, x, b) = \frac{\exp(w_k^T x + b_k)}{\sum_{j=1}^{K} \exp(w_j^T x + b_j)} = \mathrm{SoftMax}(Wx + b)$$

$$p(y_1 \dots y_N \mid x_1, \dots, x_N, W, b) = \prod_{i=1}^{N} p(y_i \mid x_i, w, b) \to \max_{w,b}$$

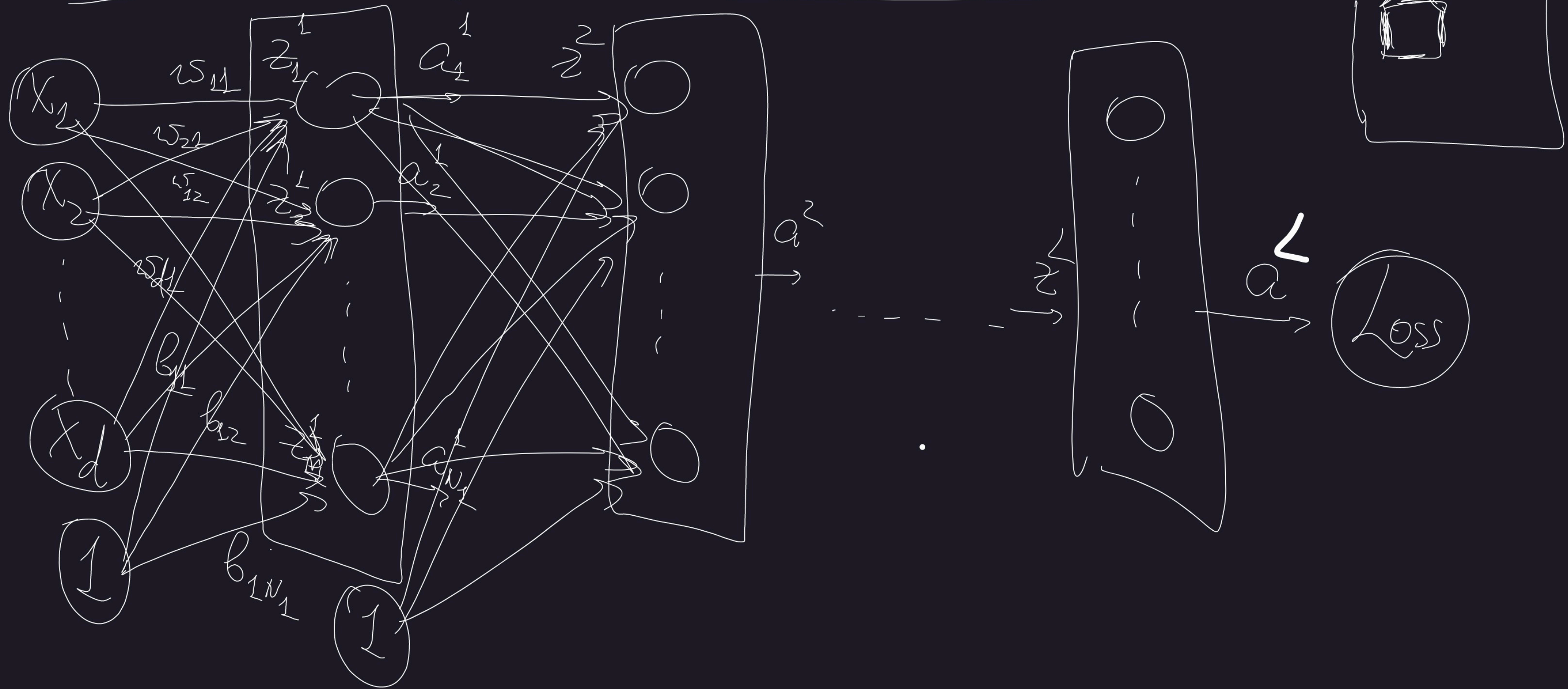
$$-\log p(y \mid X, W, b) = -\sum_{i=1}^{N} \log p(y_i \mid x_i, w, b) \to \min_{w,b}$$




$$z = w^T x + b$$

$$a = \sigma(z)$$

# Multilayer Perceptron (MLP)



$a^0 := x$

for $l = 1 \ldots L$:

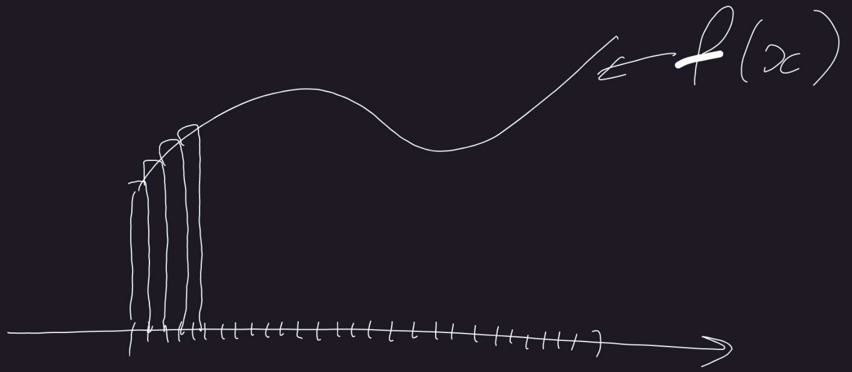
$\quad z^l = W^l a^{l-1} + b^l$

$\quad a^l = \sigma(z^l)$

output $Loss(a^L)$

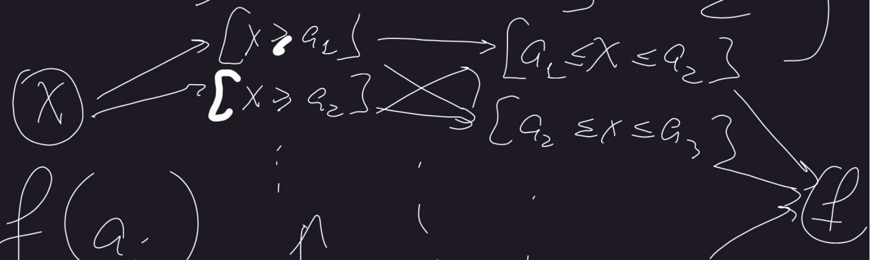
$\Theta := \{W^1, b^1, W^2, b^2, \ldots, W^L, b^L\}$

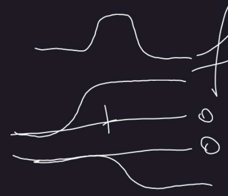
$$\sum_{i=1}^{N} Loss(y_i, a^L(x_i, \Theta)) \to \min_{\Theta}$$




$$[a_i \le X \le a_{i+1}] =$$

$$= \left[ [X \ge a_i] + [X \le a_{i+1}] \ge \frac{3}{2} \right]$$

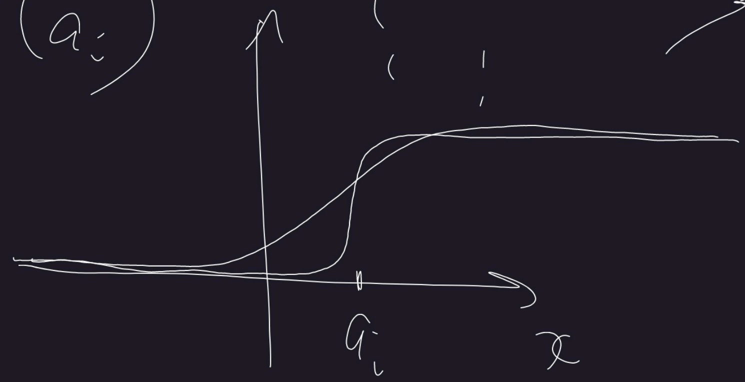


$$f(x) \approx \sum_i I[a_i \le x \le a_{i+1}]\, f(a_i)$$

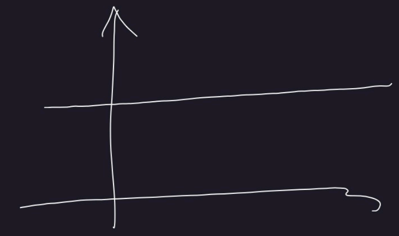
$$[X \ge a_i] \approx \sigma(k(x - a_i))$$

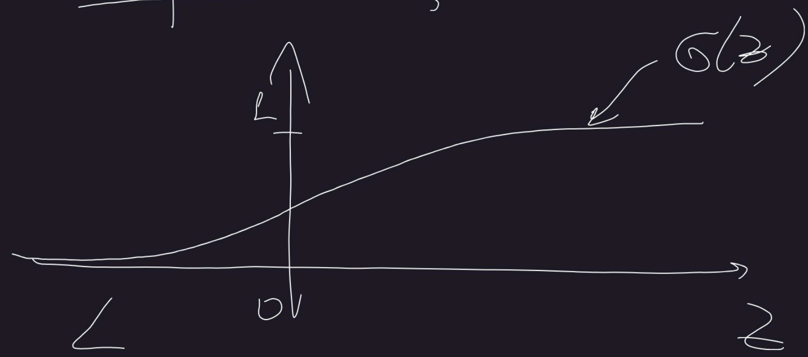
$$[X \le a_{i+1}] = 1 - [X \ge a_{i+1}] \quad k >> 1$$

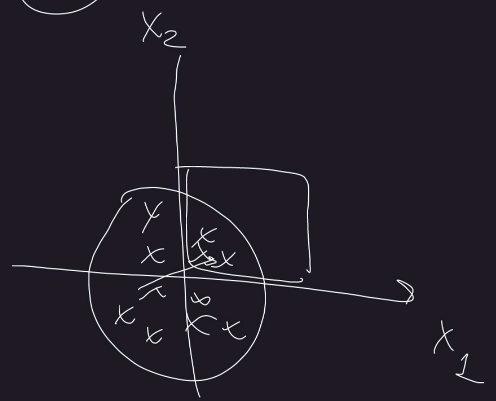


# Activation functions

(1) $a = \sigma(z) = \dfrac{1}{1+\exp(-z)}$





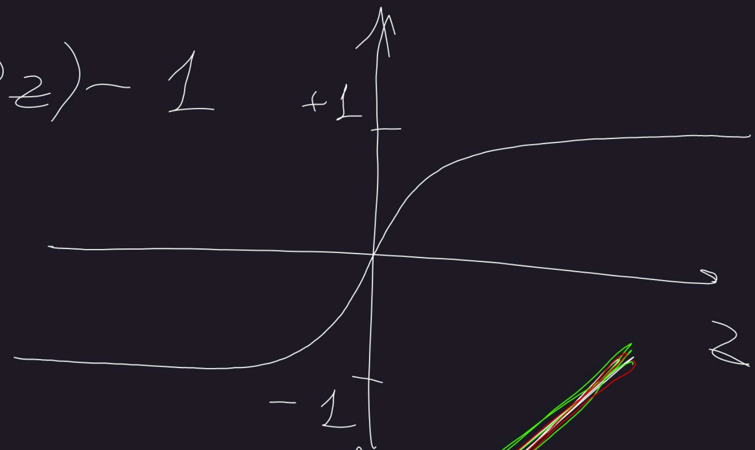




$$\nabla_{a^l} Loss = \prod_{i=l+1}^{L} (W^l)^T \frac{\partial a_i}{\partial z_i} \cdot \nabla_{a^L} Loss$$

$$\frac{\partial a_i}{\partial z_i} = \mathrm{diag}(\sigma'(z_{ij}))$$

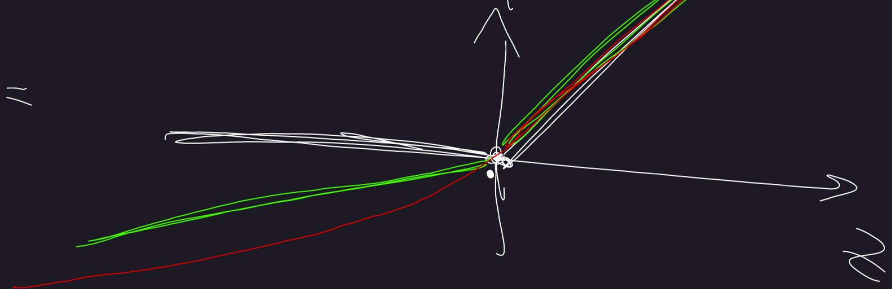
$$\|\nabla_{a^l} Loss\| \le \prod_{i=l+1}^{L} \|W^l\| \cdot \left\|\frac{\partial a_i}{\partial z_i}\right\| \cdot \|\nabla_{a^L} Loss\|$$

2. $a = \tanh(z) = 2\sigma(2z) - 1$




3. $a = \mathrm{ReLU}(z) = \max(0, z)$

4. $a = \mathrm{LeakyReLU}(z) = \begin{cases} z, & z \geq 0 \\ \alpha z, & z < 0 \end{cases} \quad \alpha \in (0, 1)$




5. $a = \mathrm{ELU}(z) = \begin{cases} z, & z \geq 0 \\ \alpha(\exp(z) - 1), & z \leq 0 \end{cases}$

# Automatic Diff.

Given: $f(x)$ Goal: $\nabla f(x)$

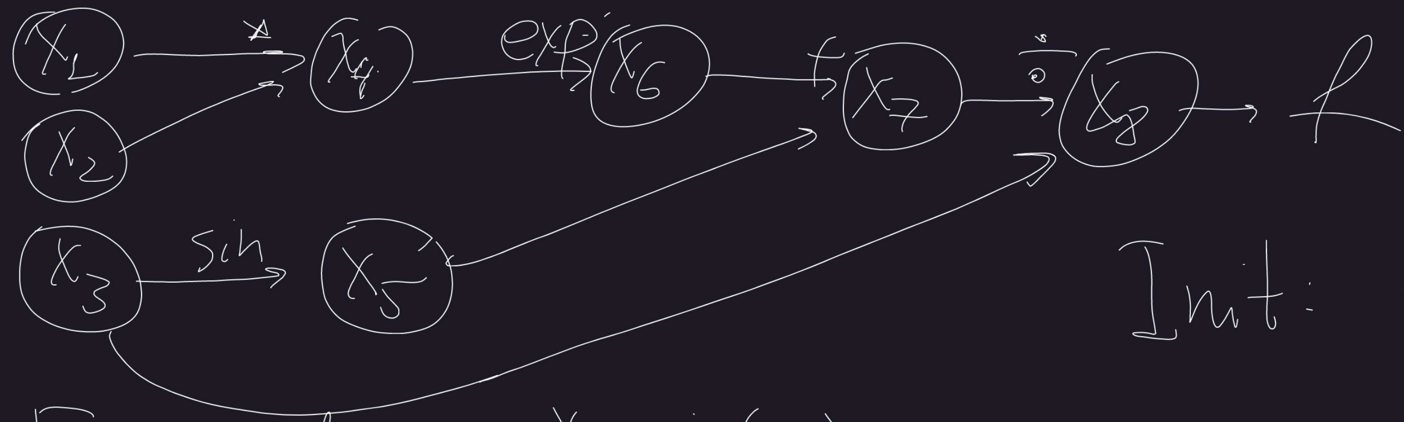
$$\frac{\partial f}{\partial x_i} \approx \frac{f(x+\varepsilon e_i) - f(x-\varepsilon e_i)}{2\varepsilon}$$

$$e_i = [0 \dots 0\,1\,0 \dots 0]$$

$i^{th}$ pos.

$$\varepsilon = \varepsilon_m^{1/3}$$

$$f(x_1, x_2, x_3) = \frac{\exp(x_1 x_2) + \sin(x_3)}{x_3}$$




Init: $\frac{\partial x_1}{\partial x_1} = 1, \quad \frac{\partial x_2}{\partial x_1} = 0, \quad \frac{\partial x_3}{\partial x_1} = 0$

**Forward**

$$\frac{\partial x_j}{\partial x_1}, \quad j = 1, 2, \ldots, 9$$

$x_5 = \sin(x_3)$

$$\frac{\partial x_5}{\partial x_1} = \cos(x_3) \cdot \frac{\partial x_3}{\partial x_1}$$

$x_4 = x_1 \cdot x_2$

$$\frac{\partial x_4}{\partial x_1} = \frac{\partial x_1}{\partial x_1} \cdot x_2 + x_1 \cdot \frac{\partial x_2}{\partial x_1}$$

# Backward diff.

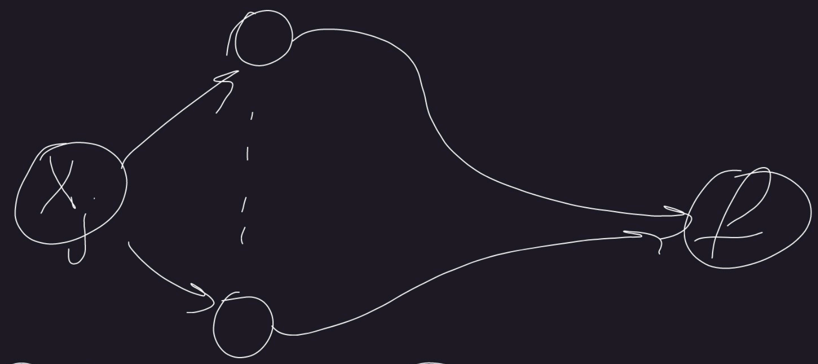
$$\frac{\partial f}{\partial x_j}, \quad j = 9, 8, \dots, 1$$

Init. $\frac{\partial f}{\partial x_8} = 1$

$$x_8 = \frac{x_7}{x_3}$$

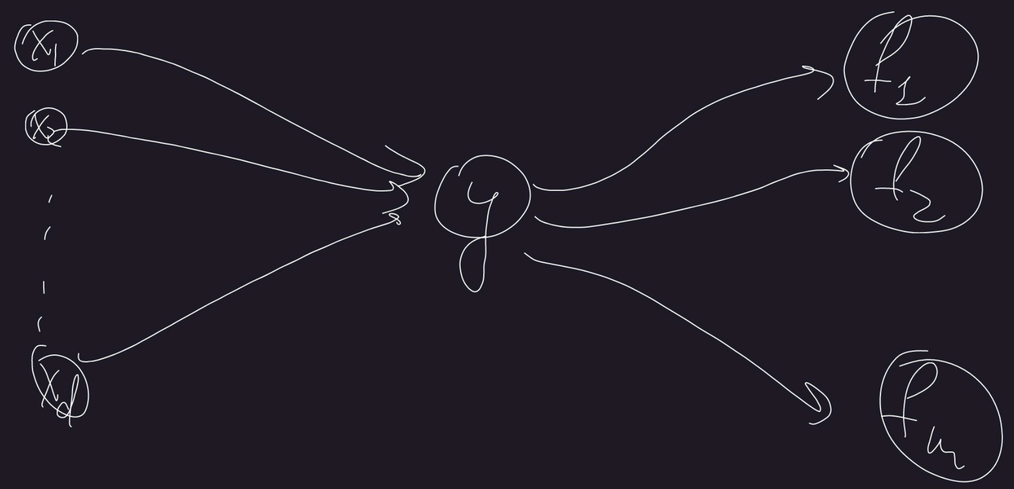
$$\frac{\partial f}{\partial x_7} = \frac{\partial f}{\partial x_8} \cdot \frac{\partial x_8}{\partial x_7}$$

$$\frac{\partial x_8}{\partial x_7} = \frac{1}{x_3}$$




$$\frac{\partial f}{\partial x_j} = \sum_{i:(j,i)\in E} \underbrace{\frac{\partial f}{\partial x_i}}_{\text{prev. it.}} \cdot \underbrace{\frac{\partial x_i}{\partial x_j}}_{\text{compute directly}}$$

# Cross mode




$$\frac{\partial f_i}{\partial x_j}, \quad i = 1 \ldots m, \quad j = 1 \ldots d$$

$$\frac{\partial y}{\partial x_j} \cdot \frac{\partial f_i}{\partial y} = \frac{\partial f_i}{\partial x_j}$$

$$f(X) = \operatorname{tr}(\underset{n\times n}{A}\,\underset{n\times n}{X^{-1}}\,\underset{n\times n}{B}) : \mathbb{R}^{n\times n} \to \mathbb{R}$$

$$\nabla f(X) - ?$$

$$\frac{\partial f}{\partial X} = \frac{\partial}{\partial X}\operatorname{tr}(AX^{-1}B) = \operatorname{tr}\Big(\underbrace{\frac{\partial}{\partial X}\underbrace{AX^{-1}B}_{n\times n}}_{4d\ \text{tensor}}\Big)$$

$$f : U \to V$$

differential

$$f(x+h) - f(x) = df(x)[h] + \bar{o}(\|h\|)$$

linear op. w.r.t. $h$

$$f : \mathbb{R} \to \mathbb{R} \qquad df = f'(x)\cdot dx$$

$$f : \mathbb{R}^n \to \mathbb{R} \qquad df = \nabla f(x)^T dx$$

$$f : \mathbb{R}^{n\times n} \to \mathbb{R} \qquad df = \operatorname{tr}(\nabla f(x)^T dX)$$

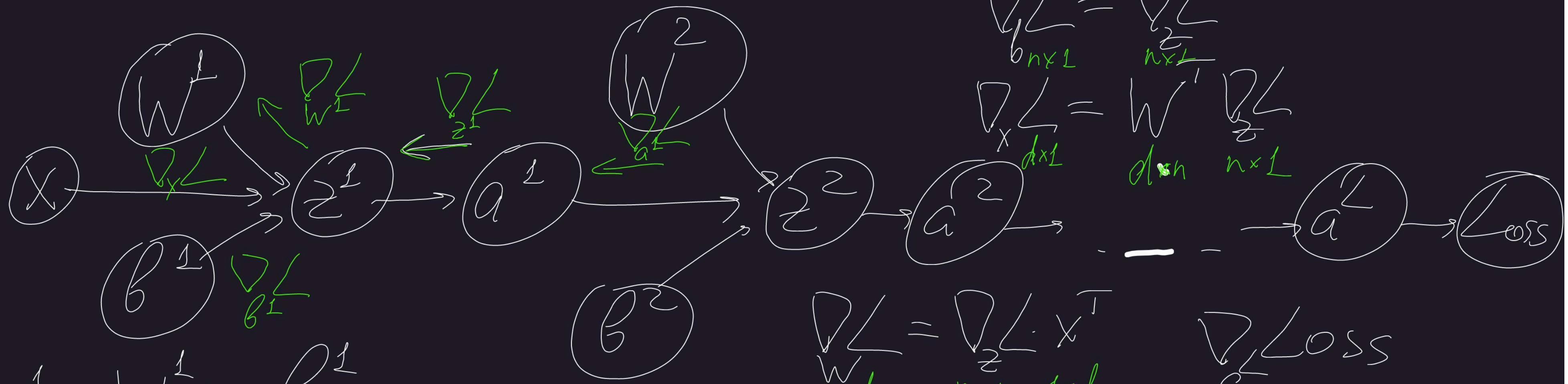
$$f(X) = tr(A X^{-1} B)$$

$$df = tr(d(A X^{-1} B)) = tr(A \underbrace{dX^{-1}}_{= -X^{-1} dX \cdot X^{-1}} B) =$$

$$= -tr(A X^{-1} dX (X^{-1} B)) =$$

$$= -tr(X^{-1} B A X^{-1} dX)$$

$$Df(X) = -X^{-T} A^{T} B^{T} X^{-T}$$




$$\underset{n\times 1}{\nabla_b L} = \underset{n\times 1}{\nabla_z L}$$

$$\underset{d\times 1}{\nabla_x L} = \underset{d\times n}{W^{1}} \underset{n\times 1}{\nabla_z L}$$

$$\underset{n\times d}{\nabla_W L} = \underset{n\times 1}{\nabla_z L} \cdot \underset{1\times d}{x^T}$$

$$\nabla_{a^L} Loss$$

$$\underset{n\times 1}{z^1} = \underset{n\times d}{W^1} \cdot \underset{d\times 1}{x} + \underset{n\times 1}{b^1}$$

$$a^1 = g(z^1)$$

$$dL = \nabla_{z^1} L^T dz^1 = tr(\nabla_{W^1} L^T dW^1) + \nabla_x L^T dx + \nabla_{b^1} L^T db^1$$

$$dz = d(Wx + b) = dW \cdot x + W \cdot dx + db$$

$$dL = \nabla_z L^T (dW \cdot x + W dx + db) = tr(\nabla_z L^T dW x) + \nabla_z L^T W dx + \nabla_z L^T db$$

$$a = g(z)$$

$$da = \frac{\partial a}{\partial z} \cdot dz$$

$$\overset{\Downarrow}{\mathrm{diag}(g'(z))}$$

$$dL = \nabla_a L^T da = \underbrace{\nabla_a L^T \frac{\partial a}{\partial z}}_{\nabla_z L^T} \cdot dz$$

$$\nabla_z L = \left(\frac{\partial a}{\partial z}\right)^T \cdot \nabla_a L$$