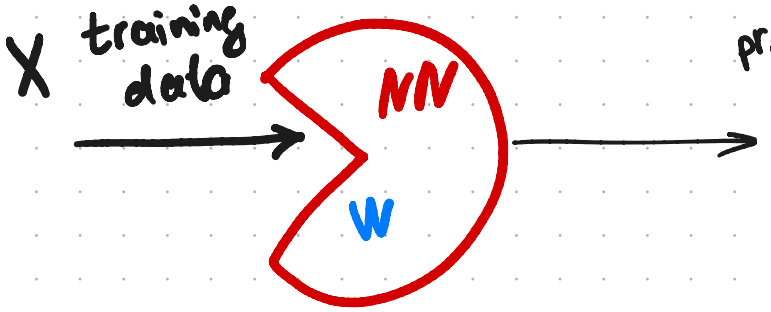


W - model parameters (weights)

$$Loss(w, X) \simeq \frac{1}{N} \sum_{i=1}^{N} \left(y_{pred}^{i}(w, X) - y_{true}^{i} \right)^2 \rightarrow$$

P is extremely large

$\rightarrow \min_{w \in \mathbb{R}^p}$

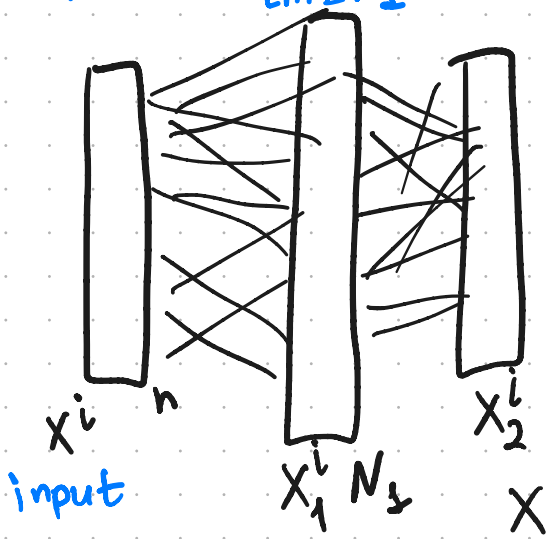
Adam or SGD or SGD with momentum

$w, \nabla_w L, V, S$ — $4p$

$w \in \mathbb{R}^p$, $\nabla_w L \in \mathbb{R}^p$ — $2p$

$w, \nabla_w L, V$ — $3p$

$y_{pred}(w, X)$ - NEURAL NETWORK



$$x_2^i = W_2 \cdot W_1 \cdot x^i$$

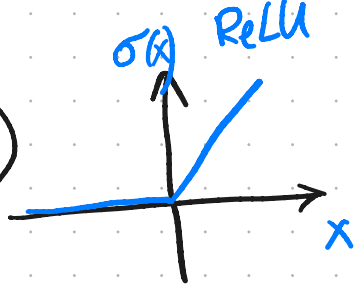
$N_2 \times 1$ $N_2 \times N_1$ $N_1 \times n$ $n \times 1$

GPT-3, transformers and etc.

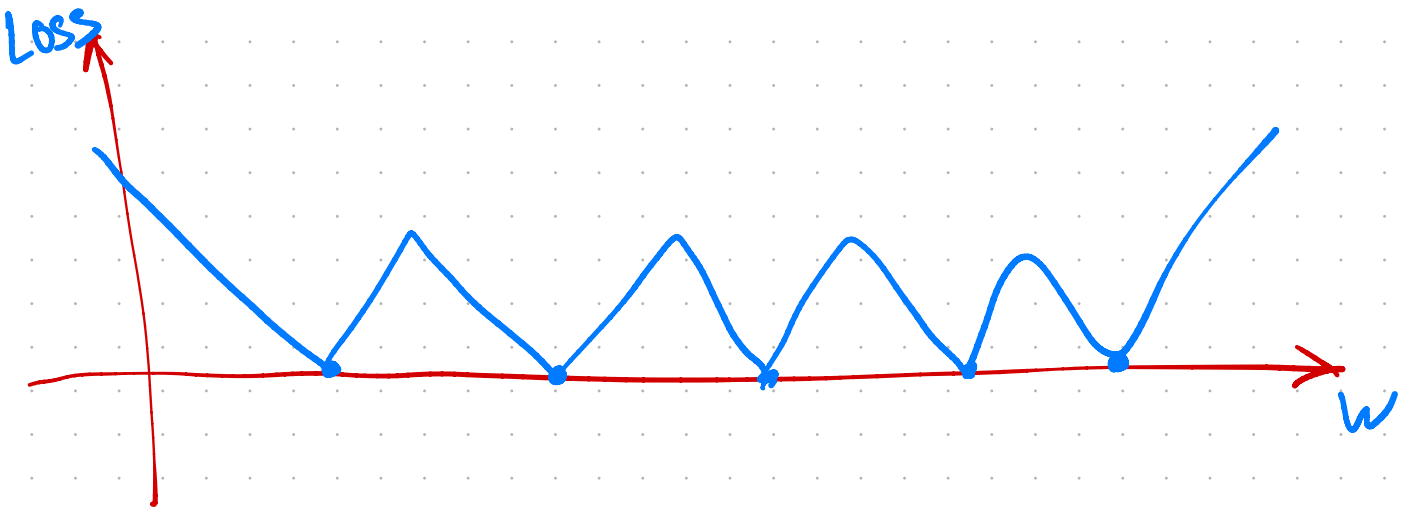
1 GPU? (255 years)

activation function

$$x_1^i = \sigma_1 \left(W_1 \cdot x^i + b_1 \right)$$



PROBLEM IS NOT CONVEX.



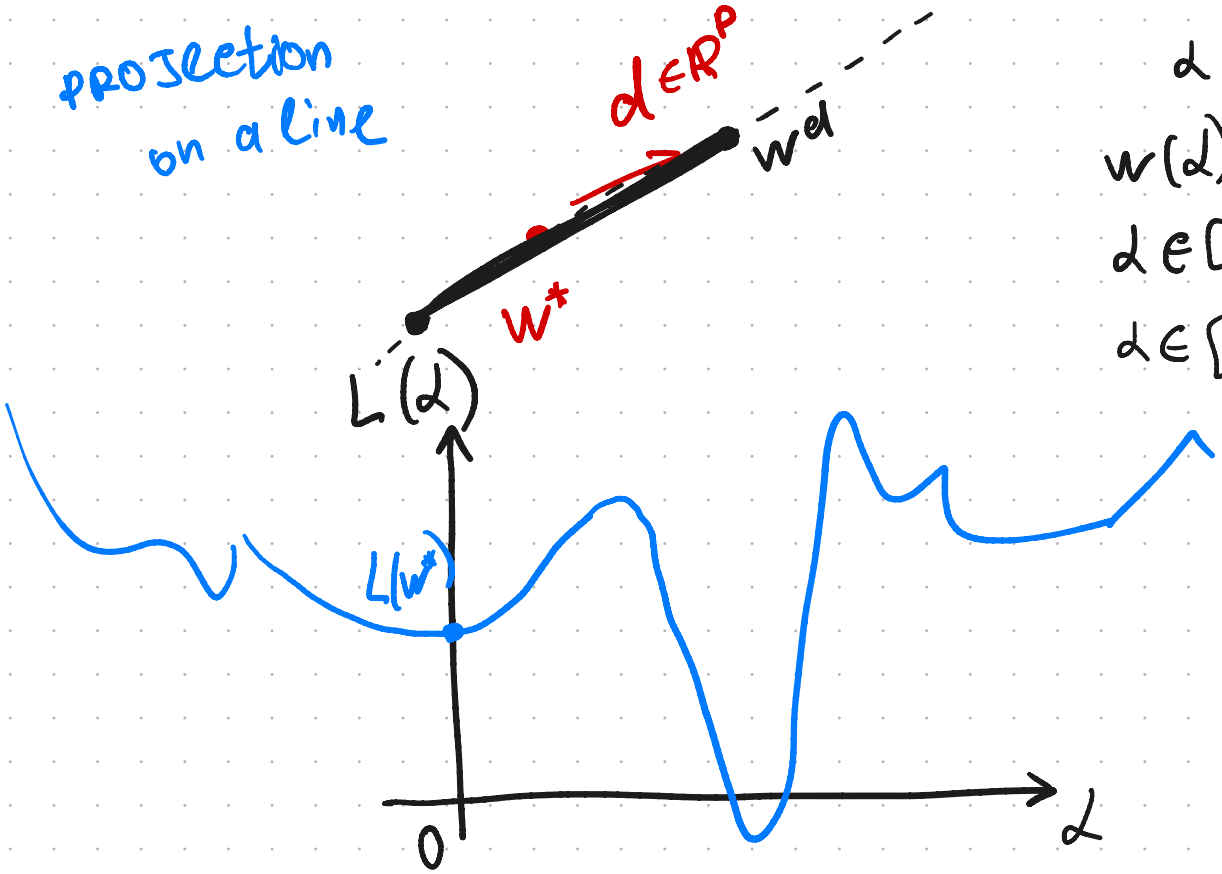
$L(w) \to \min_{w \in \mathbb{R}^p}$

w^*

$w(\alpha) = w^* + \alpha \cdot d$

$w(\alpha) = (1-\alpha)w^* + \alpha \cdot w^d$

projection on a line



$\alpha = 0$

$w(\alpha) = w^*$

$\alpha \in [0, 1]$

$\alpha \in [-1, 1]$

w_1, w_2

$w^*, w, w_1, w_2 \in \mathbb{R}^p$

w^*

$$w(\alpha, \beta) = w^* + \alpha \cdot w_1 + \beta w_2$$

$\alpha \in [-1, 1]$

$\beta \in [-1, 1]$

$\alpha, \beta \in \mathbb{R}$

$L(w(\alpha, \beta)) = \dots$