Back propagation & Optimization

<u>Plan:</u>

Review

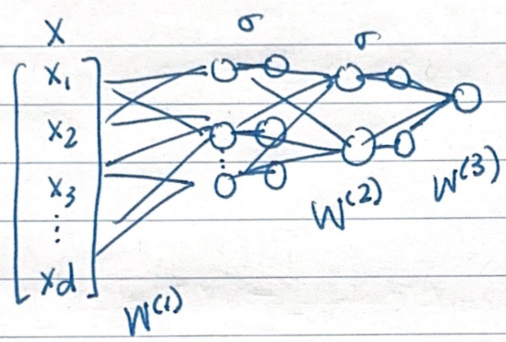
Backprop

Stochastic Gradient Descent

<u>Logistics</u>

Zoom + earbud

Review

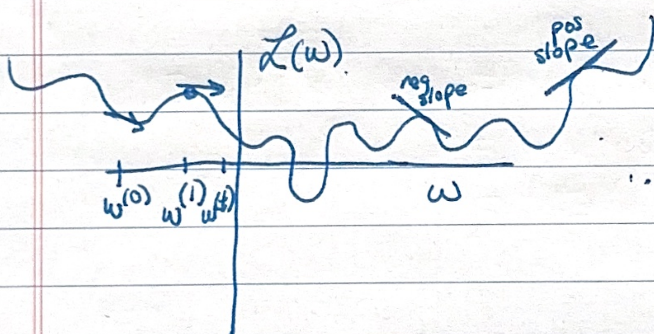


<u>Activations</u>

- ReLU
- sigmoid
- hyper tan
- ⋮

<u>Layers</u>

- FC
- Conv
- Residual
- Attention



$$w^{(t+1)} = w^{(t)} - \alpha \nabla_w \mathcal{L}(w)$$

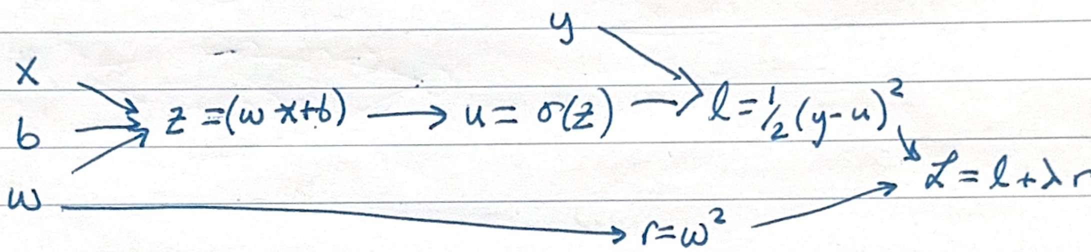
Backprop

$$\mathcal{L}(w,b) = \frac{1}{2}\left(y - \sigma(w\cdot x+b)\right)^2 + \lambda w^2$$ "regularization" $\qquad w, b \in \mathbb{R}$

$$\frac{\partial \mathcal{L}}{\partial w} = (y-\sigma(w\cdot x+b))\cdot\sigma'(wx+b)\,x \quad + 2\lambda w$$

$$\frac{\partial \mathcal{L}}{\partial b} = (y-\sigma(w\cdot x+b))\cdot -\sigma'(wx+b)\,\cancel{x}$$

⊖ complicated ⊖ Redundant



Forward

for $i \in \{1, \ldots, N\}$:

compute v_i as function of Parents(v_i)

Backward $i \in \{N, \ldots, 1\}$:

compute $$\frac{\partial \mathcal{L}}{\partial v_i} = \sum_{j \in \text{children}(v_i)} \frac{\partial \mathcal{L}}{\partial v_j}\cdot\frac{\partial v_j}{\partial v_i}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial z}\overbrace{\frac{\partial z}{\partial b}}^{1} \leftarrow \frac{\partial \mathcal{L}}{\partial z} = \frac{\partial \mathcal{L}}{\partial u}\frac{\partial u}{\partial z} \leftarrow \frac{\partial \mathcal{L}}{\partial u} = \frac{\partial \mathcal{L}}{\partial l}\frac{\partial l}{\partial u} \leftarrow \frac{\partial \mathcal{L}}{\partial l} = 1 \leftarrow \mathcal{L}$$

$$\frac{\partial \mathcal{L}}{\partial u} = 1\cdot(y-u)\cdot -1$$

$$\frac{\partial \mathcal{L}}{\partial z} = (u-y)\cdot\sigma'(z)$$

$$\frac{\partial \mathcal{L}}{\partial r} = \lambda$$

$$\frac{\partial \mathcal{L}}{\partial w} = x$$

$$\frac{\partial r}{\partial w} = 2w$$

Autograd

$$\frac{\partial \mathcal{L}}{\partial v_i} = \sum_{j \in \text{children}(i)} \frac{\partial \mathcal{L}}{\partial v_j} \cdot \frac{\partial v_j}{\partial v_i}$$

$$= \begin{bmatrix} \cdots \frac{\partial \mathcal{L}}{\partial v_j} \cdots & \end{bmatrix} \begin{bmatrix} \vdots \\ \frac{\partial v_j}{\partial v_i} \\ \vdots \end{bmatrix}$$

$$\begin{bmatrix} \frac{\partial \mathcal{L}}{\partial v_1} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial v_n} \end{bmatrix} = \begin{bmatrix} \cdots \frac{\partial \mathcal{L}}{\partial v_1} \cdots \\ \\ \cdots \frac{\partial \mathcal{L}}{\partial v_j} \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ \frac{\partial v_j}{\partial v_1} \\ \vdots \\ \frac{\partial v_i}{\partial v_j} \end{bmatrix}$$

matrix multiplication
↳ forward!
↳ Backward!

GPU* really good at this
- lots of cores ⇒ parallel
- lots of memory

*valuable so load in batches

GD : ~~Lw~~ $\frac{1}{n} \sum_{i=1}^{n} l_i(x^{(i)}, y^{(i)})$

vs.

Stochastic GD: $S \subseteq [n]$

$$\frac{1}{|S|} \sum_{i \in S}^{n} l_i(x^{(i)}, y^{(i)})$$

SGD Variants

$$w^{(t+1)} = w^{(t)} - \alpha \nabla_w \mathcal{L}(w^{(t)})$$

Momentum!

- ↳ get out of local optima
- ↳ "average" out noise

$$v^{(t+1)} = \sqrt{\beta}\, v^{(t)} + \sqrt{(1-\beta)}\, \nabla_w \mathcal{L}(w^{(t)})$$
$$w^{(t+1)} = w^{(t)} - \alpha v^{(t)}$$

Adaptive!

- ↳ adjust step with progress

$$s^{(t+1)} = \sqrt{\beta}\, s^{(t)} + \sqrt{(1-\beta)}\, \| \nabla_w \mathcal{L}(w^{(t)}) \|_2^2$$
$$w^{(t+1)} = w^{(t)} - \frac{\alpha}{\sqrt{s^{(t+1)}} + \epsilon} \nabla_w \mathcal{L}(w^{(t)})$$

AdaM!!

- ↳ both!!

$$w^{(t+1)} = w^{(t)} - \frac{\alpha}{\sqrt{s^{(t+1)}} + \epsilon} v^{(t)}$$

Initialization

$$f(x) \approx \sigma(W^{(L)}\sigma(W^{(L-1)} \cdots \sigma(W^{(1)} x))\cdots)$$

↳ Random so no symmetry

↳ Var so values don't explode or vanish

$$z^{(l+1)} = W^{(l)} z^{(l)} \qquad W^{l} \in \mathbb{R}^{n_{in} \times n_{out}}$$

$$z_i^{(l+1)} = \sum_{j=1}^{n_{in}} W_{ij}^{(l)} z_j^{(l)} \qquad Var(W_{ij}^{(l)}) = \sigma^2$$

$$Var(z_i^{(l+1)}) \overset{indep}{=} \sum_{j=1}^{n_{in}} \sigma^2 \, var(z_j^{(l)}) = n_{in} \sigma^2 \, var(z_j^{(l)})$$

$$\sigma^2 n_{in} \approx 1$$

Next layer: $\sigma^2 n_{out} \approx 1$ so $\sigma^2 = \frac{1}{2n_{in} + 2n_{out}}$

Generalization

Test vs training data

- bottlenecks
- early stopping
- weight decay
- dropout
- transfer learning