Back Prop / Gradients

$$\nabla_x = W^T \frac{\partial \mathcal{L}}{\partial \ldots}$$
$$\nabla_w = \frac{\partial \mathcal{L}}{\partial \ldots} x^T$$

Max/Relu · vector: When $Mx=\vec{v}$ on Θ

$$\text{diag}\left(\mathbb{1}(\tfrac{v}{?}) + \mathbb{1}(\tfrac{v}{?}), \ldots, \text{"n"}\right)$$

Don't forget 'Relu's · use names:

$$\cos \text{sim} = \frac{a^T b}{\|a\|\|b\|}$$

quotient Rule see handout

$$\frac{\partial \log |X|}{\partial X} = (X^T)^{-1}$$
$$\frac{\partial \mathcal{L}_i}{\partial X} = W^T \frac{\partial \mathcal{L}_i}{\partial (WX)}$$
$$\frac{\partial \mathcal{L}_i}{\partial W} = \frac{\partial \mathcal{L}_i}{\partial (Wx)} x^T$$
$$\frac{\partial \mathcal{L}_i}{\partial X} = 2 \frac{\partial \mathcal{L}_i}{\partial (XX^T)} X$$

Derivative with respect to a tensor
If $XY = Z \longrightarrow \mathcal{L}$, and we have an upstream derivative $\partial \mathcal{L}/\partial Z$:

$$\frac{\partial \mathcal{L}}{\partial X} = \frac{\partial Z}{\partial X}\frac{\partial \mathcal{L}}{\partial Z} = \frac{\partial \mathcal{L}}{\partial Z}Y^T \quad (Y^T \text{ is on the right})$$
$$\frac{\partial \mathcal{L}}{\partial Y} = \frac{\partial Z}{\partial Y}\frac{\partial \mathcal{L}}{\partial Z} = X^T\frac{\partial \mathcal{L}}{\partial Z} \quad (X^T \text{ is on the left})$$

$Wx = c \longrightarrow \ell$, with $W \in \mathcal{R}^{h \times m}, x \in \mathcal{R}^m, c \in \mathcal{R}^h$.

$$\underbrace{\frac{\partial \ell}{\partial x}}_{m \times 1} = \underbrace{\frac{\partial c}{\partial x}}_{m \times h}\underbrace{\frac{\partial \ell}{\partial c}}_{h \times 1} = \underbrace{W^T}_{m \times h}\underbrace{\frac{\partial \ell}{\partial c}}_{h \times 1} \quad (W^T \text{ is on the left})$$
$$\underbrace{\frac{\partial \ell}{\partial W}}_{h \times m} = \underbrace{\frac{\partial c}{\partial W}}_{h \times m \times h}\underbrace{\frac{\partial \ell}{\partial c}}_{h \times 1} = \underbrace{\frac{\partial \ell}{\partial c}}_{h \times 1}\underbrace{x^T}_{1 \times m} \quad (x^T \text{ is on the right})$$

Regularization

L2: Min dist or make smaller — Feb 2019
L1: Sparsity — aBe 2019

'remember λ in front'

Too much reg can underfit

Activations

	Saturates flat	Differentiable continuous	Zig Zag always positive	Identity $f(0)=0$ $f'(0)=1$
tanh	yes ($\pm$)	yes	no	yes
ReLU	~no	no	yes	no
PReLU	no	no	no	no
ELU	yes (-)	no	no	no
Sigmoid	yes ($\pm$)	yes	yes	no
Exp	yes (-)	yes	yes	no

KNN / Linear

KNN- slow test, high var, low bias

Linear - High bias, low var

Log-likelihood:

$$\mathcal{L} = \prod_i^n P(x_i | \theta)$$
$$\log \mathcal{L} = \sum \log P(x|\theta)$$

$$\mathcal{L} = \prod_{i=1}^{n} p(x^i, y^i | \theta)$$
$$\log \mathcal{L} = \sum \log p(x,y|\theta)$$
$$= p(y) \, p(x|y)$$

Binomial
$$P(|\theta) = \theta^{n_1}(1-\theta)^{n_2}, \quad n_1 + n_2 = n$$

Softmax! equiprob

$$\text{argmax}_\theta = \prod P(x|\theta) \, p(y|x,\theta)$$

Convolution

Standard Output size:
$$\left(\frac{w - w_f + 2\,pad}{\text{stride}} + 1, \text{ " "}\right) / \text{filter}$$

Pooled Output!
$$\left(\frac{w - w_p}{\text{stride}} + 1, \frac{h - h_p}{\text{stride}} + 1\right) / \text{filter}$$
— leaves depth same

Parameters:
$$(h_f \cdot w_f \cdot d) \cdot \text{\# filters (or output depth)}$$
— \#/neuron

Operations: (2 × filter) — 2mul Flops

Output (filter + (filter - 1))
  multi   sum

pool   Output × (filter - 1)

Expectation + Estimators

Estimator Bias: $E[\hat{\theta}] - \theta$
Var: $E[(\hat{\theta} - E[\hat{\theta}])^2]$

Linear: $E[x+y] = E[x] + E[y]$

E of non-RV is value itself

E of mean 0 is 0