* activation (tranformation)
  initialization "pre-training"
  optimization
  regularization L2, early stopping

* activation

$$S_j^{(l)} = \sum_{i=0}^{d^{(l-1)}} W_{ij}^{(l)} x_i^{(l-1)}$$

$$x_j^{(l)} = \phi_j^{(l)}\left(S_j^{(l)}\right)$$

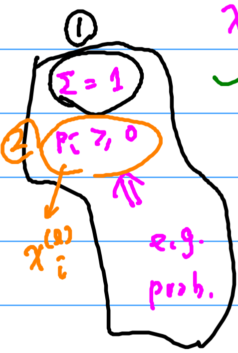
scalar

⇓

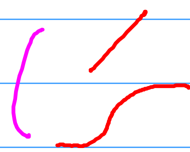
vector activation

$$\underset{\text{joint}}{x^{(l)}} = \underset{\text{property}}{\phi}\left(S^{(l)}\right)$$

real values



$\phi_j^{(L)}$ Last layer : desired output



regression

logistic bin. classification

$\phi_j^{(l)}$ hidden layer : "soft perc."

tanh

multi class

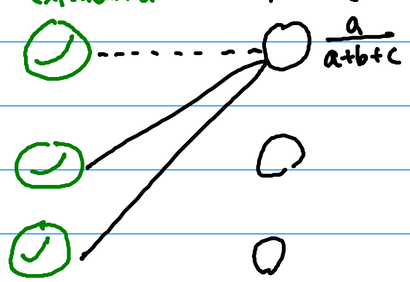
* softmax activation

output often

$$\phi(S) = \left(\frac{e^{S_1}}{\Sigma e^{S_i}}, \frac{e^{S_2}}{\Sigma e^{S_i}}, \frac{e^{S_3}}{\Sigma e^{S_i}}, \ldots, \frac{e^{S_d}}{\Sigma e^{S_i}} \right)$$

≈ 0 ≈ 1

exponential ratio (normalized)



"general" backprop

$$\boxed{\text{activation}}$$

$$s_j^{(\ell)} = \sum_{i=0}^{d^{(\ell-1)}} w_{ij}^{(\ell)} \boxed{x_i^{(\ell-1)}}$$

$$x_j^{(\ell)} = \phi_j^{(\ell)}\left(s_j^{(\ell)}\right)$$

* bottleneck of being deep:

$$\phi_j^{(\ell)}$$

$$\nabla_{ij}^{(\ell)} = \delta_j^{(\ell)} \, x_i^{(\ell-1)}$$

$$\delta_j^{(\ell)} = \frac{\partial e}{\partial s_j^{(\ell)}} = \sum_k \delta_k^{(\ell+1)} w_{jk}^{(\ell+1)} \phi'\left(s_j^{(\ell)}\right)$$

$$= \sum_k \sum_m \delta_m^{(\ell+2)} \left(w_{km}^{(\ell+2)} w_{jk}^{(\ell+1)}\right) \phi'\left(s_k^{(\ell+1)}\right) \phi'\left(s_j^{(\ell)}\right)$$

$$\boxed{\delta_j^{(1)}} = \left(\underset{\text{many}}{w}\right) \cdot \boxed{\underset{\text{many}}{\phi'}} \, \delta_j^{(L)}$$

* traditional NNet

$\phi(s) = \tanh(s)$ $\quad \phi'(s) \in (0, 1)$

$x_j^{(\ell)} \in (-1, 1)$

① $w_{ij}^{(\ell)}$ large → s large

$\phi(s) \approx 1$ (or -1) $\quad$ saturation $\quad \phi'(s) \approx 0$

② $\boxed{w_{ij}^{(\ell)} \text{ small}}$

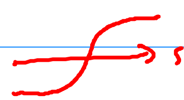
→ $s_j^{(\ell)}$ small $\quad \phi'$ "shrink"

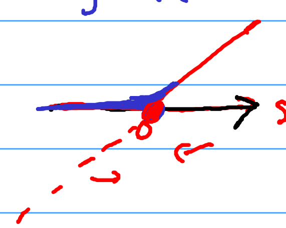
earlier layer small gradient

vanishing gradient

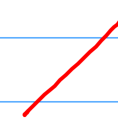
* modern deep learning

$\phi(s) = \max(s, 0)$





$x_j^{(\ell)} \in [0, \infty)$ $\quad$ non-linear, monotonic



rectified linear unit (ReLU) $\quad \nabla = \delta \, x$

$$\phi'(s) = \begin{cases} \boxed{1 : s \geq 0} \\ 0 : s < 0 \end{cases} \qquad \boxed{\phi' \ \phi' \ \phi'} \quad \nu$$

dead

- (o) – less issues on gradient vanishing
- (o) – more efficient operations (no exp, log ...)
- (o) – sparsity
- (△) – not fully differentiable ϕ'