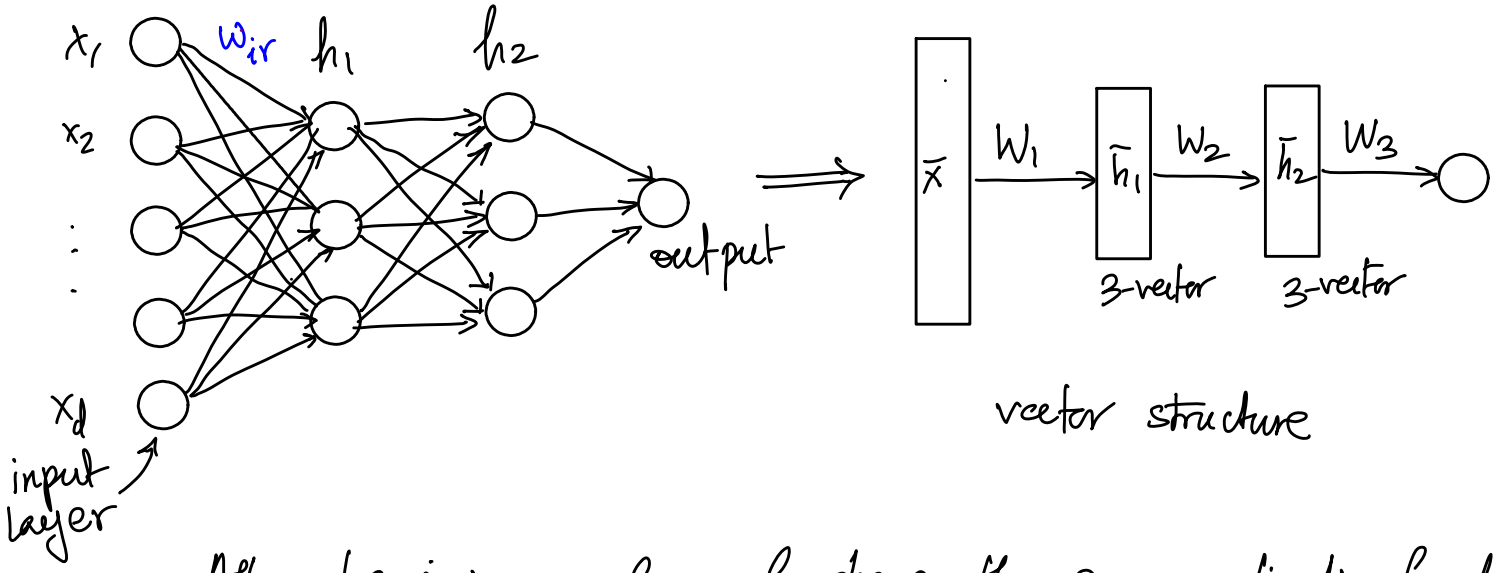
MATH 565: Lecture 28 (04/23/2026)

Today: * vector case of NNs

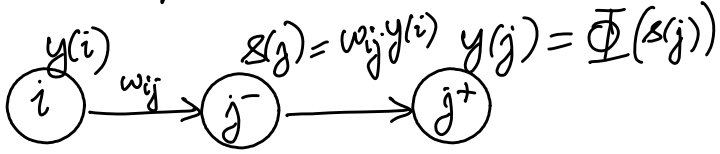
Feedforward NNs as Vector Structures

Recall the previous illustration:



All nodes j in a layer ℓ share the same activation function Φ, receive inputs from all nodes in layer $(\ell-1)$, and pass output to each node in layer $(\ell+1)$.

Recall the decoupled formulation:



Generalizing to vector settings, we get

$i \to j^-$: $\bar{s}^{(\ell)} = W^{(\ell)} \bar{y}^{(\ell-1)}$

→ vector of weighted sums

$j^- \to j^+$: elementwise activation

$\bar{y}^{(\ell)} = \underline{\Phi}(\bar{s}^{(\ell)})$

Note: In some settings, $\bar{y}^{(\ell)}$ is taken as $\text{diag}(\underline{\Phi}'(\bar{s}))$ (↗ diagonal matrix) so that it has same dimensions as associated Jacobians. But we'll stick with the vector form.

Sensitivity Recursion

$$\bar{s}(j^-, t) = \bar{s}(j^+, t) \odot \bar{\Phi}'(\bar{z})$$

→ Hadamard or elementwise product

It is standard notation to use $\bar{\delta}^-$ to denote $\bar{s}(j^-, t)$ or $\bar{\delta}^{-(\ell)}$ for the sensitivities at level ℓ. Hence, the recursion is given as

$$\bar{\delta}^{-(\ell)} = \bar{\delta}^{+(\ell)} \odot \bar{\Phi}'(\bar{z}^{(\ell)}).$$

Plugging in the recursive expression for $\bar{\delta}^{+(\ell)}$ gives us the backward pass equation as

$$\bar{\delta}_{prev} = W^T \bar{\delta}^- \quad \text{or, with levels,}$$

$$\bar{\delta}^{+(\ell-1)} = W^{(\ell)T} \bar{\delta}^{-(\ell)}.$$

We next formally prove these recursive relations for vector back propagation.

Theorem 15 (Vector backpropagation) Let $G_c = (N, \mathcal{E})$ be an N-layer feedforward NN with $\bar{y}^{(0)} = \bar{x}$ and $\bar{s}^{(\ell)} = W^{(\ell)} \bar{y}^{(\ell-1)}$, $\bar{y}^{(\ell)} = \bar{\Phi}(\bar{s}^{(\ell)})$, with $\bar{\Phi}$ applied elementwise, for $\ell = 1, \dots, N$, and loss function $J(\bar{y}^{(N)})$. Let $\bar{\delta}^{-(\ell)} = \frac{\partial J}{\partial \bar{s}^{(\ell)}}$, $\bar{\delta}^{+(\ell)} = \frac{\partial J}{\partial \bar{y}^{(\ell)}}$ be the pre- and post-activation sensitivity vectors at layer ℓ. Then the following relations hold for $\ell = 1, \dots, N$:

$$\bar{\delta}^{-(\ell)} = \bar{\delta}^{+(\ell)} \odot \bar{\Phi}'(\bar{s}^{(\ell)}) \quad \text{——— (1)}$$

$$\bar{\delta}^{+(\ell)} = W^{(\ell+1)T} \bar{\delta}^{-(\ell+1)} \quad \text{——— (2)}$$

$$\frac{\partial J}{\partial W^{(\ell)}} = \bar{\delta}^{-(\ell)} \bar{y}^{(\ell-1)T} \quad \text{——— (3)}$$

with initialization $\bar{\delta}^{+(N)} = \nabla_{\bar{y}^{(N)}} J$.

Proof We apply chain rule backwards from layer N. At the output layer,

$$\bar{\delta}^{+(N)} = \nabla_{\bar{y}^{(N)}} J \quad \text{by definition.}$$

$$\Rightarrow \quad \bar{\delta}^{-(N)} = \frac{\partial J}{\partial \bar{s}^{(N)}} = \frac{\partial J}{\partial \bar{y}^{(N)}} \odot \frac{\partial \bar{y}^{(N)}}{\partial \bar{s}^{(N)}}$$

$$= \bar{\delta}^{+(N)} \odot \bar{\Phi}'(\bar{s}^{(N)}),$$

giving (1) for $\ell = N$.

Induction

Assume $\bar{\delta}^{-(l+1)}$ is known for some l in $1 \le l < N$. As J depends on $\bar{y}^{(l)}$ only through $\bar{s}^{(l+1)} = W^{(l+1)} \bar{y}^{(l)}$, chain rule gives

$$\bar{\delta}^{+(l)} = \frac{\partial J}{\partial \bar{y}^{(l)}} = \left(\frac{\partial \bar{s}^{(l+1)}}{\partial \bar{y}^{(l)}}\right)^T \frac{\partial J}{\partial \bar{s}^{(l+1)}}$$

($W^{(l+1)}$)

$$= W^{(l+1)T} \bar{\delta}^{-(l+1)}, \text{ giving (2).}$$

Applying entrywise activation derivative then gives

$$\bar{\delta}^{-(l)} = \frac{\partial J}{\partial \bar{s}^{(l)}} = \frac{\partial J}{\partial \bar{y}^{(l)}} \odot \bar{\Phi}'(\bar{s}^{(l)})$$

$$\Rightarrow \quad \bar{\delta}^{-(l)} = \bar{\delta}^{+(l)} \odot \bar{\Phi}'(\bar{s}^{(l)}),$$

giving (1) at layer l.

Since J depends on $W^{(l)}$ only through $\bar{s}^{(l)} = W^{(l)} \bar{y}^{(l-1)}$, differentiating entrywise gives

$$\frac{\partial J}{\partial W_{jk}^{(l)}} = \sum_i \frac{\partial J}{\partial s_i^{(l)}} \cdot \frac{\partial s_i^{(l)}}{\partial W_{jk}^{(l)}} = \bar{\delta}_j^{-(l)} \cdot y_k^{(l-1)}, \text{ since}$$

$$\frac{\partial s_i^{(l)}}{\partial W_{jk}^{(l)}} = y_k^{(l-1)} \cdot \delta_{ij}.$$ Assembling over all (j,k) entries gives $\bar{\delta}^{-(l)} \cdot \bar{y}^{(l-1)T}$, proving (3).

□

We next present a result that is directly implied by the above recursions and characterizes some limitations on trainability of NNs just based on their architectures.

Theorem 16 (Exponential decay of sensitivities) Suppose $\|\dot{\Phi}'\|_\infty \le \gamma$ and $\|W^{(l)}\|_2 \le \sigma$ for all layers $l=1,\dots,N$.

Then $\|\delta^{-(l)}\|_2 \le (\sigma\gamma)^{N-l} \|\delta^{-(N)}\|_2$

Spectral norm: $\|W\|_2 = \sup_{\|\bar{x}\|_2=1} \|W\bar{x}\|_2$ is the largest singular value of W. We note that $\|AB\|_2 \le \|A\|_2\|B\|_2$, called the submultiplicity of spectral norm.

Proof (1) and (2) $\Rightarrow$

$$\delta^{-(l)} = \left(W^{(l+1)^T}\delta^{-(l+1)}\right) \odot \dot{\Phi}'(s^{(l)})$$

$$\Rightarrow \|\delta^{-(l)}\|_2 \le \|W^{(l+1)^T}\delta^{-(l+1)}\|_2 \cdot \|\dot{\Phi}'(s^{(l)})\|_\infty \le \gamma$$

using the standard Hadamard bound inequality:

$$\|\bar{u}\odot\bar{v}\|_2 \le \|\bar{u}\|_2\|\bar{v}\|_\infty$$

$$\le \|W^{(l+1)}\|_2 \|\delta^{-(l+1)}\|_2 \cdot \gamma$$

$\le \sigma$ by submultiplicativity of spectral norm

$$\Rightarrow \|\delta^{-(l)}\|_2 \le \sigma\gamma\|\delta^{-(l+1)}\|_2.$$

Applying this relation (or contraction) inductively gives

$$\|\delta^{-(l)}\|_2 \le (\sigma\gamma)^{N-l}\|\delta^{-(N)}\|_2.$$

□

So, if $\sigma\gamma < 1$, then $\|\bar{s}^{(\ell)}\|_2$ decays exponentially as we go from layer N toward the input layer ($\ell = 1$).

Hence, (3) $\Rightarrow \frac{\partial J}{\partial W^{(\ell)}}$ decays exponentially for early layers, rendering it hard or impossible to train the NN with GD.

Theorem 16 suggests the following design choices for NNs.

* ReLU activation: $\bar{\Phi}'(s) = 1\,(s > 0) \Rightarrow \gamma = 1$. So $\bar{\Phi}$ does not contribute to the contraction of sensitivities. In contrast, sigmoid activation has $\bar{\Phi}' \le \frac{1}{4}$ $\Rightarrow \gamma \le \frac{1}{4} \Rightarrow$ decay factor is $\left(\frac{1}{4}\right)^{N-\ell}$

* Choosing $W^{(\ell)}$ such that $\|W^{(\ell)}\|_2 = 1$ keeps $\sigma = 1$ (so, no contraction or amplification). Combined with ReLU activation, we get $\sigma\gamma = 1$, and the sensitivities are not forced to decay just by the NN architecture alone.

* Residual NNs with skip connections could bypass this issue: