The Role of Sparsity in Learning With Overparameterized Deep Neural Networks

Rahul Parhi
UCSD ECE

Conference on Mathematics of Data Science

21 October 2024

A Brief History of Neural Networks and AI

1943: McCulloch and Pitts had the vision to introduce artificial intelligence to the world.

BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 5, 1943

A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S. MCCULLOCH AND WALTER PITTS

FROM THE UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE, DEPARTMENT OF PSYCHIATRY AT THE ILLINOIS NEUROPSYCHIATRIC INSTITUTE, AND THE UNIVERSITY OF CHICAGO

1958: Rosenblatt implemented the first perceptron for learning.

Psychological Review
Vol. 65, No. 6, 1958

THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN [1]

F. ROSENBLATT

Cornell Aeronautical Laboratory

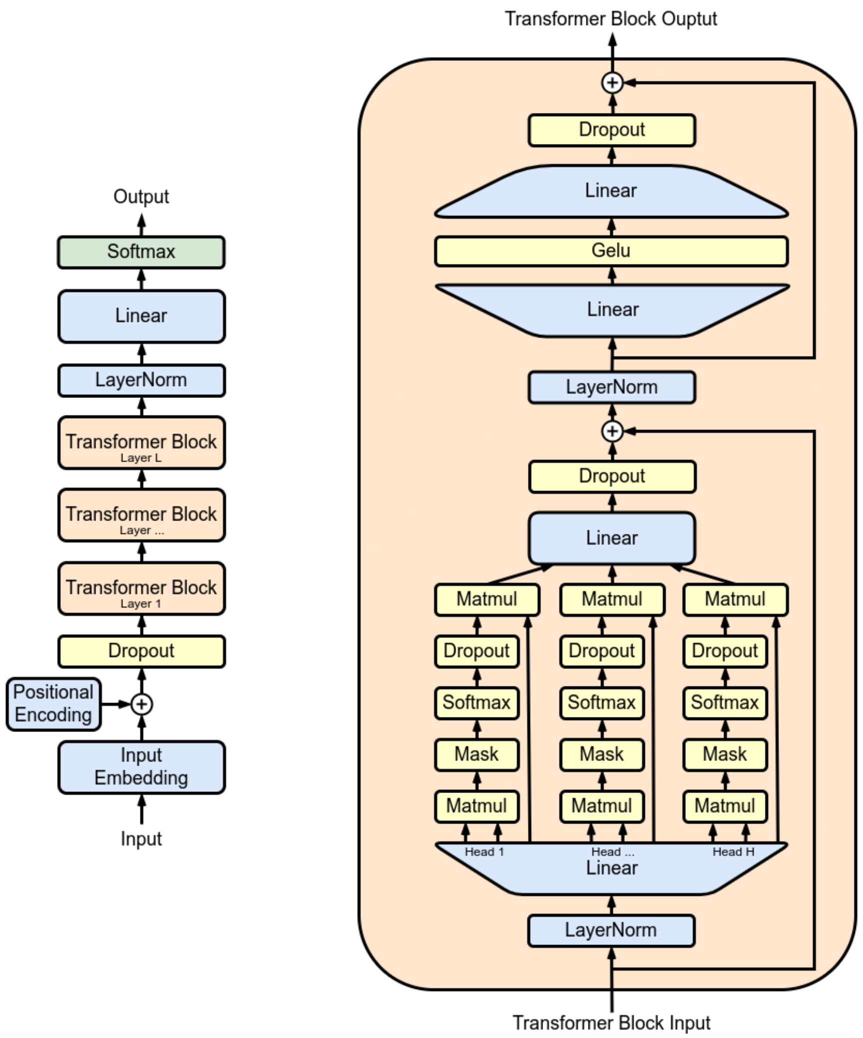
1986: Rumelhart, Hinton, and Williams studied backpropagation for training multilayer perceptrons.

Learning representations by back-propagating errors

David E. Rumelhart*, Geoffrey E. Hinton† & Ronald J. Williams*

* Institute for Cognitive Science, C-015, University of California, San Diego, La Jolla, California 92093, USA
† Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Philadelphia 15213, USA

The World Is Now Based on Neural Networks



Large language models (LLMs) like generative pre-trained transformers (GPT) have taken the world by storm.

- ChatGPT
- Claude

Do we even understand why neural networks work?

[PDF] Improving language understanding by generative pre-training

A Radford, K Narasimhan, T Salimans, I Sutskever

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document …

Save Cite Cited by 6469 Related articles

Magnetic Resonance Imaging (MRI)

Accelerating MRI scans is one of the principal outstanding problems in the MRI research community.

- Early approaches were based on **compressed sensing**.

Magnetic Resonance in Medicine 58:1182–1195 (2007)

Sparse MRI: The Application of Compressed Sensing for Rapid MR Imaging

Michael Lustig,[1]* David Donoho,[2] and John M. Pauly[1]

$\Longrightarrow$ Theoretical guarantees of **stability**.

Candès et al. (2006)
Donoho (2006)

- Modern approaches are based on **deep learning** and massive amounts of **data**.

2306 IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 40, NO. 9, SEPTEMBER 2021

Results of the 2020 fastMRI Challenge for Machine Learning MR Image Reconstruction

Matthew J. Muckley, *Member, IEEE*, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, *Member, IEEE*, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, Simon Arberet, Dominik Nickel, Zaccharie Ramzi, *Student Member, IEEE*, Philippe Ciuciu, *Senior Member, IEEE*, Jean-Luc Starck, Jonas Teuwen, Dimitrios Karkalousos, Chaoping Zhang, Anuroop Sriram, Zhengnan Huang, Nafissa Yakubova, Yvonne W. Lui, and Florian Knoll, *Member, IEEE*

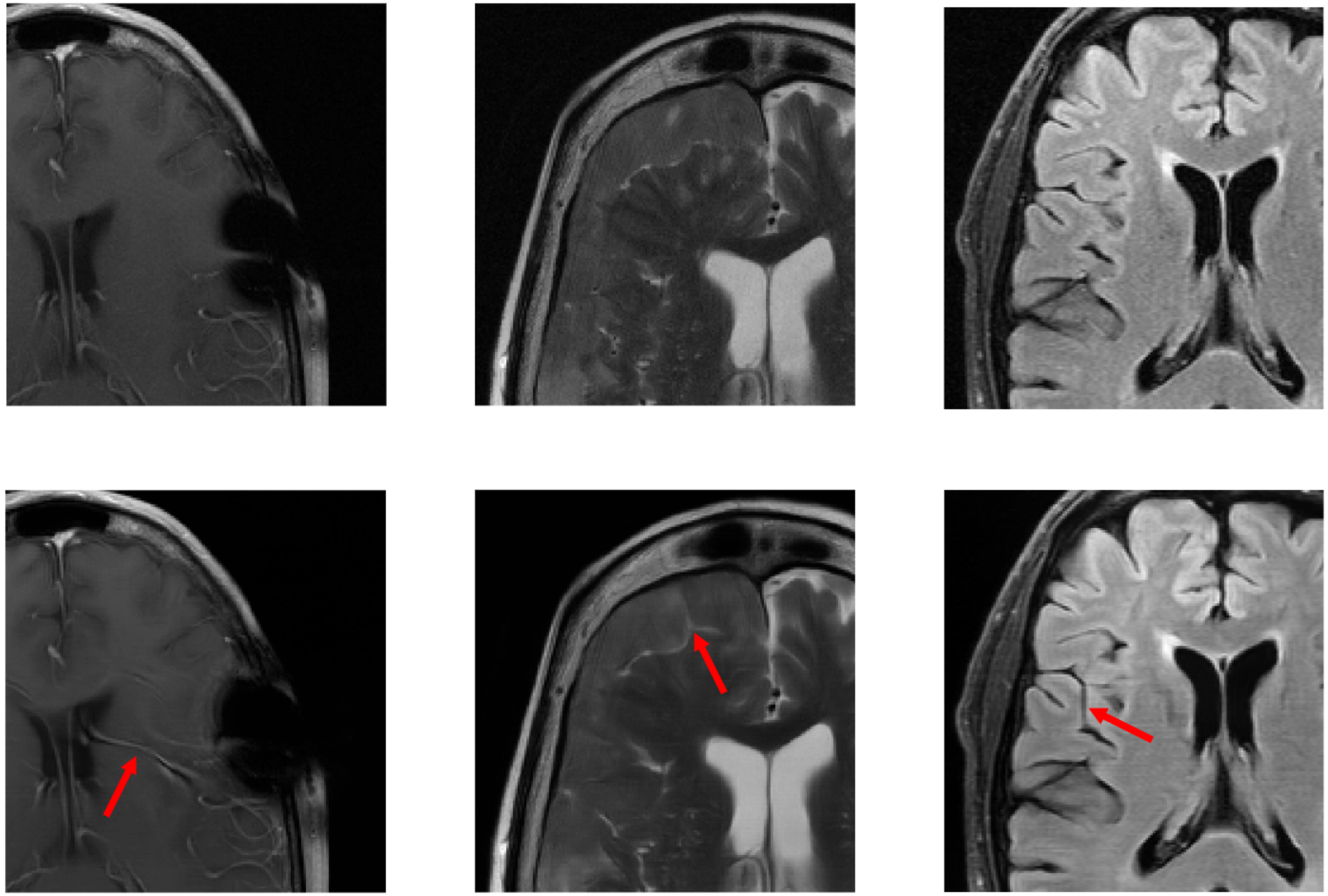
$\Longrightarrow$ Almost no theoretical guarantees.

Can we trust deep-learning-based methods?

Results of the 2020 fastMRI Challenge

Ground Truth

DNN-Based Reconstruction

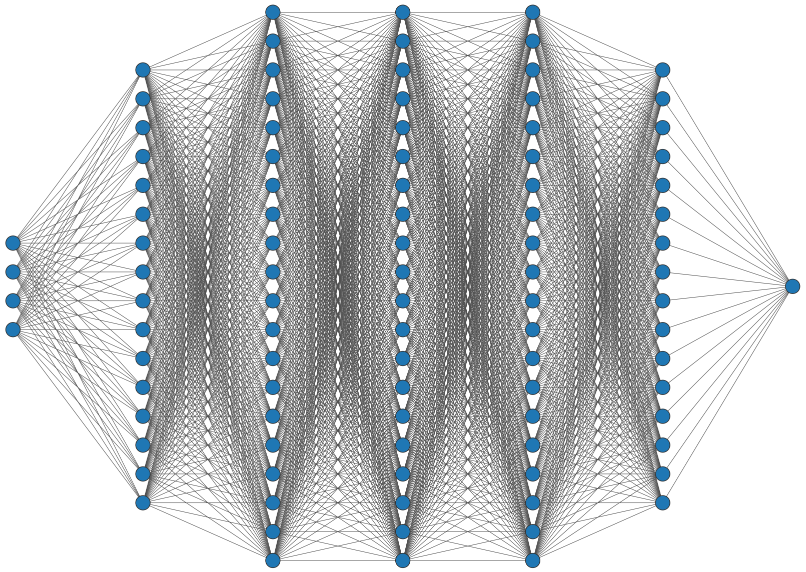


AI-generated hallucinations identified by radiologists as **false** vessels.

Muckley et al. (2021, IEEE Transactions on Medical Imaging)

Today's Talk

Understanding **analytic properties** of **trained** neural networks.

$\boldsymbol{x}$  $f_{\boldsymbol{\theta}}(\boldsymbol{x})$

parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^P$ of neural network **weights**

Neural network training problem for the data $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^P} \underbrace{\sum_{n=1}^{N} \mathcal{L}(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n))}_{\text{data fidelity}} + \underbrace{\frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2}_{\text{regularization}}$$

← Tikhonov regularization "weight decay"

We will be **agnostic** to the optimization algorithm.

Joint Work With...


Joe Shenouda


Kangwook Lee


Rob Nowak

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

JMLR

Variation Spaces for Multi-Output Neural Networks: Insights on Multi-Task Learning and Network Compression

Joseph Shenouda, Rahul Parhi, Kangwook Lee, Robert D. Nowak; 25(231):1–40, 2024.

Weight Decay in Neural Network Training

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^P} \underbrace{\sum_{n=1}^{N} \mathcal{L}(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n))}_{\mathscr{L}(\boldsymbol{\theta})} + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

weight decay objective

Gradient descent update on θ_i

weight decay

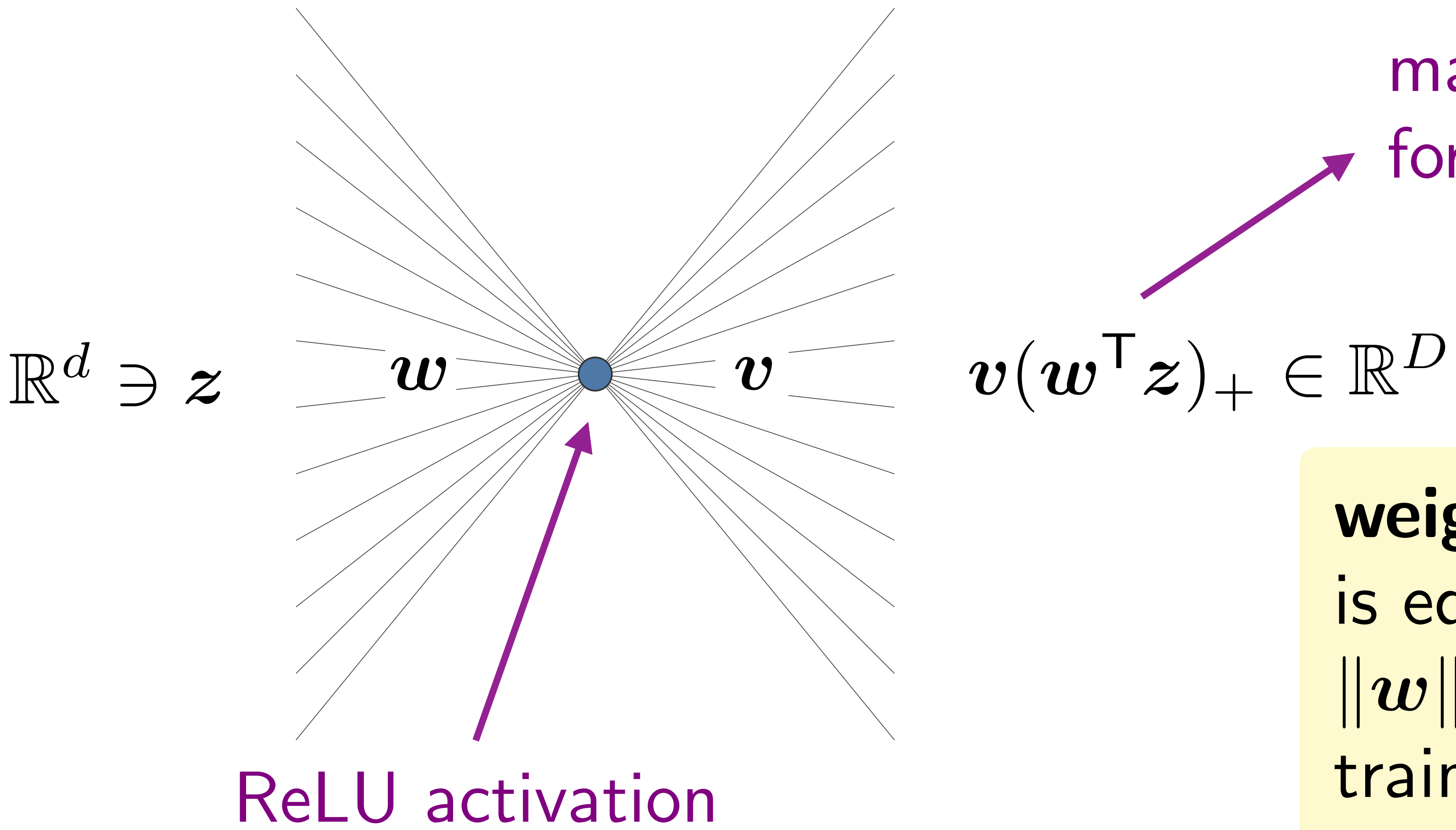
$$\theta_i^{t+1} = \theta_i^t - \tau \left(\frac{\partial \mathscr{L}}{\partial \theta_i}\bigg|_{\theta_i=\theta_i^t} + \lambda\theta_i^t \right) = \theta_i^t - \tau \frac{\partial \mathscr{L}}{\partial \theta_i}\bigg|_{\theta_i=\theta_i^t} - \tau\lambda\theta_i^t$$

step size
"learning rate"

GD update on $\mathscr{L}$

Hanson and Pratt (1988, NeurIPS)
Krogh and Hertz (1990, NeurIPS)

Neural Balance in Deep Neural Networks



weight decay in training is equivalent to adding $\|\boldsymbol{w}\|_2^2 + \|\boldsymbol{v}\|_2^2$ to the training objective

Neural Balance Theorem

If a DNN is trained with weight decay, then the 2-norms of the input and output weights to each ReLU neuron must be **balanced**.

$$\|\boldsymbol{w}\|_2 = \|\boldsymbol{v}\|_2$$

Yang, Zhang, Shenouda, Papailiopoulos, Lee, and Nowak (2022)
P. and Nowak (2023)

Neural Balance

The ReLU activation is **homogeneous**

$$\boldsymbol{v}(\boldsymbol{w}^\top \boldsymbol{z})_+ = \gamma^{-1}\boldsymbol{v}(\gamma \boldsymbol{w}^\top \boldsymbol{z})_+, \quad \text{for any } \gamma > 0.$$

At a global minimizer of the weight decay objective, $\|\boldsymbol{v}\|_2 = \|\boldsymbol{w}\|_2$.

Proof. The solution to

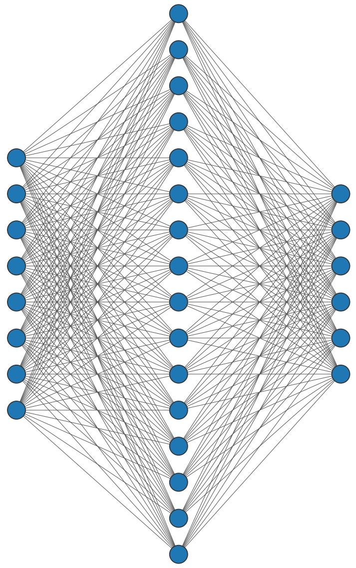
$$\min_{\gamma > 0} \|\gamma^{-1}\boldsymbol{v}\|_2 + \|\gamma \boldsymbol{w}\|_2$$

is $\gamma = \sqrt{\|\boldsymbol{v}\|_2 / \|\boldsymbol{w}\|_2}$. □

At a global minimizer, $\dfrac{\|\boldsymbol{v}\|_2^2 + \|\boldsymbol{w}\|_2^2}{2} = \|\boldsymbol{v}\|_2 \|\boldsymbol{w}\|_2$.

Grandvalet (1998, ICANN)
Neyshabur et al. (2015, ICLR Workshop)

Secret Sparsity of Weight Decay



$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{v}_k (\boldsymbol{w}_k^\top \boldsymbol{x})_+$$

$$\boldsymbol{\theta} = \{(\boldsymbol{w}_k, \boldsymbol{v}_k)\}_{k=1}^{K}$$

weight decay

$$\min_{\boldsymbol{\theta}=\{(\boldsymbol{w}_k,\boldsymbol{v}_k)\}_{k=1}^{K}} \sum_{n=1}^{N} \mathcal{L}(\boldsymbol{y}_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) + \frac{\lambda}{2} \sum_{k=1}^{K} \|\boldsymbol{v}_k\|_2^2 + \|\boldsymbol{w}_k\|_2^2$$

path-norm

$$\min_{\boldsymbol{\theta}=\{(\boldsymbol{w}_k,\boldsymbol{v}_k)\}_{k=1}^{K}} \sum_{n=1}^{N} \mathcal{L}(\boldsymbol{y}_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) + \lambda \sum_{k=1}^{K} \|\boldsymbol{v}_k\|_2 \|\boldsymbol{w}_k\|_2$$

multitask lasso

$$\min_{\substack{\boldsymbol{\theta}=\{(\boldsymbol{w}_k,\boldsymbol{v}_k)\}_{k=1}^{K} \\ \|\boldsymbol{w}_k\|_2=1}} \sum_{n=1}^{N} \mathcal{L}(\boldsymbol{y}_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) + \lambda \sum_{k=1}^{K} \|\boldsymbol{v}_k\|_2$$

Rebalancing

Path-Norm and Neural Banach Spaces

$$\mathring{\mathcal{V}} = \left\{ f(\boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{v}_k (\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x})_+ \,:\, \boldsymbol{v}_k \in \mathbb{R}^D, \boldsymbol{w}_k \in \mathbb{R}^d, K \in \mathbb{N} \right\}$$

finite-width **vector-valued** networks

The path-norm is a **valid norm** on $\mathring{\mathcal{V}}$:

$$\|f\|_{\mathcal{V}} = \sum_{k=1}^{K} \|\boldsymbol{v}_k\|_2 \|\boldsymbol{w}_k\|_2$$

The "completion" of $\mathring{\mathcal{V}}$ (in an appropriate sense) is a Banach space. It is the Banach space $\mathcal{V}$ of all functions of the form

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}} (\boldsymbol{w}^\mathsf{T} \boldsymbol{x})_+ \, \mathrm{d}\boldsymbol{\nu}(\boldsymbol{w}).$$

vector-valued measure

"output weights"

Barron (1993, IEEE TIT)
Bach (2017, JMLR)
Ongie et al. (2020, ICLR)
Shenouda, **P.**, Lee, and Nowak (2024, JMLR)

Path-Norm and Vector-Valued Measures

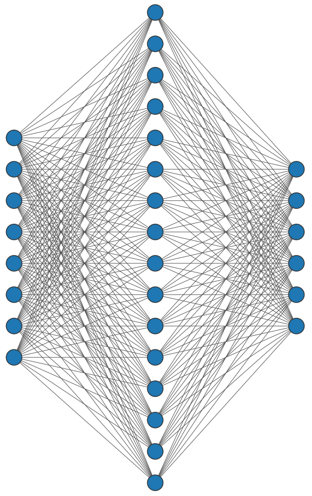
$$f \in \mathcal{V}, \quad f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}} (\boldsymbol{w}^\mathsf{T}\boldsymbol{x})_+ \,\mathrm{d}\boldsymbol{\nu}(\boldsymbol{w}), \quad \|f\|_{\mathcal{V}}$$

The measure $\boldsymbol{\nu} \in \mathcal{M}(\mathbb{R}^d; \mathbb{R}^D)$ is **vector-valued**.

$$\boldsymbol{\nu} = \begin{bmatrix} \nu_1 \\ \vdots \\ \nu_D \end{bmatrix}$$

$$\|f\|_{\mathcal{V}} = \|\boldsymbol{\nu}\|_{2,\mathcal{M}} := \sup_{\substack{\mathbb{S}^{d-1} = \bigcup_{i=1}^n A_i \\ n \in \mathbb{N}}} \sum_{i=1}^n \|\boldsymbol{\nu}(A_i)\|_2$$

$$= \sup_{\substack{\mathbb{S}^{d-1} = \bigcup_{i=1}^n A_i \\ n \in \mathbb{N}}} \sum_{i=1}^n \left(\sum_{j=1}^D |\nu_j(A_i)|^2 \right)^{1/2}$$



$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^K \boldsymbol{v}_k (\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x})_+ \implies \|f_{\boldsymbol{\theta}}\|_{\mathcal{V}} = \sum_{k=1}^K \|\boldsymbol{v}_k\|_2 \|\boldsymbol{w}_k\|_2$$

$\mathcal{V}$ is a vector-valued variation space

Shenouda, **P.**, Lee, and Nowak (2024, JMLR)

A Representer Theorem

Theorem (Shenouda, P., Lee, and Nowak 2024)

For any data set $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$ and lower semicontinuous $\mathcal{L}(\cdot,\cdot)$, there exists a solution to

$$\min_{f \in \mathcal{V}} \sum_{n=1}^{N} \mathcal{L}(\boldsymbol{y}_n, f(\boldsymbol{x}_n)) + \lambda \|f\|_{\mathcal{V}}, \quad \lambda > 0,$$

that admits a representation of the form

$$f_{\mathrm{ReLU}}(\boldsymbol{x}) = \sum_{k=1}^{K} v_k (\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x})_+ \quad \boxed{K < N^2}.$$

sparse solution

The bound is **independent** of the input and output dimensions.

Carathéodory's theorem would predict a bound of $ND + 1$.

Shenouda, **P.**, Lee, and Nowak (2024, JMLR)

Weight Decay Promotes Neuron Sharing

$$\min_{f \in \mathcal{V}} \left(J(f) := \sum_{n=1}^{N} \mathcal{L}(\boldsymbol{y}_n, f(\boldsymbol{x}_n)) + \lambda \|f\|_{\mathcal{V}} \right)$$

$\mathcal{V}$-norm regularization
$\Longleftrightarrow$
path-norm regularization
$\Longleftrightarrow$
weight decay

Neuron Sharing Theorem (Shenouda, P., Lee and Nowak 2024)

Consider a vector-valued neural network (with unique input weights)

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{v}_k (\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x})_+.$$

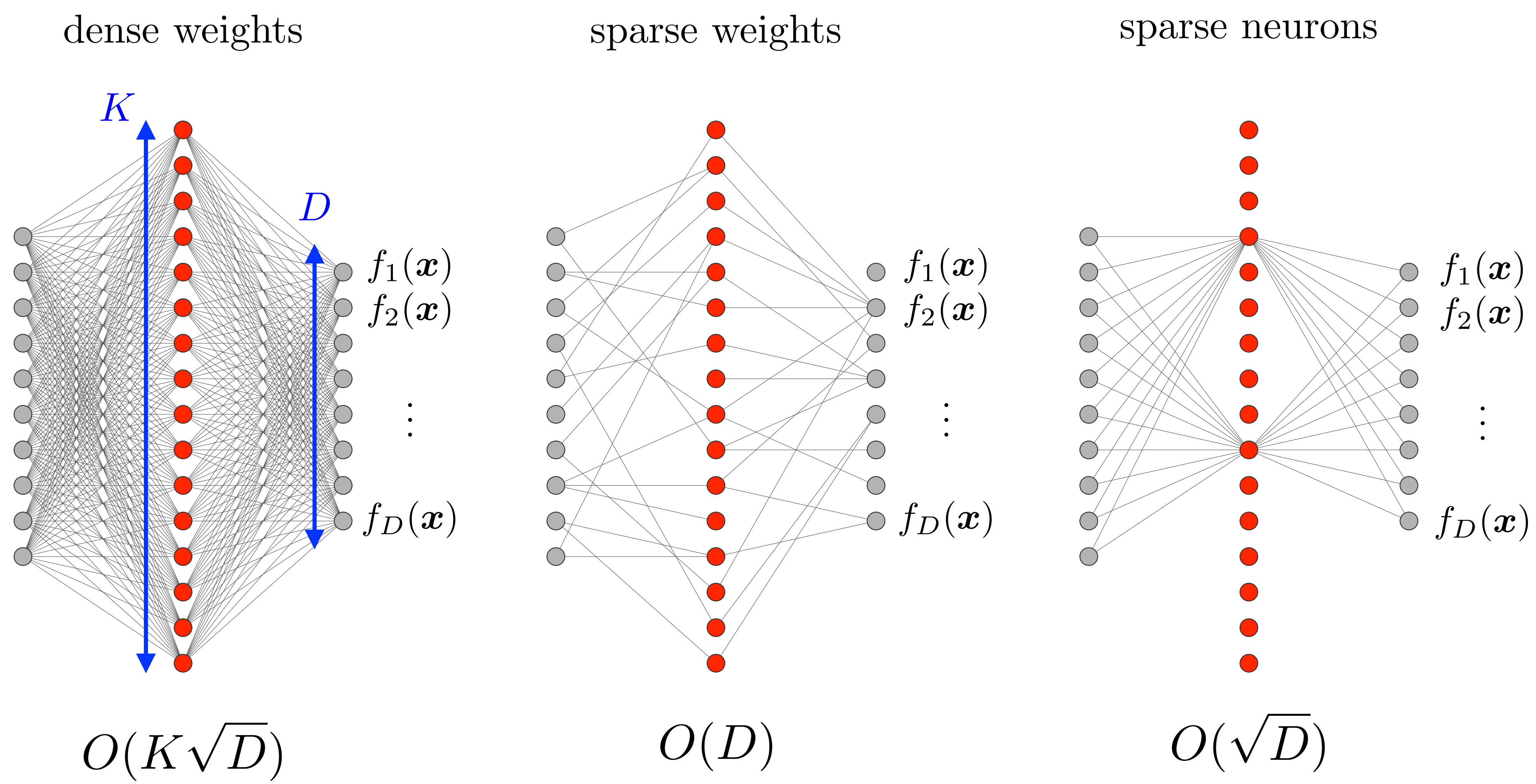
There exists $\delta > 0$ such that, if $\angle(\boldsymbol{w}_1, \boldsymbol{w}_2) < \delta$, then the neural network that *shares neurons* has a strictly smaller objective value. That is,

$$\widetilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}) - \boldsymbol{v}_1(\boldsymbol{w}_1^\mathsf{T} \boldsymbol{x}) + \widetilde{\boldsymbol{v}}_1(\boldsymbol{w}_2^\mathsf{T} \boldsymbol{x})$$

satisfies $J(\tilde{f}) < J(f)$.

Shenouda, **P.**, Lee, and Nowak (2024, JMLR)

The Structured Sparsity of Weight Decay

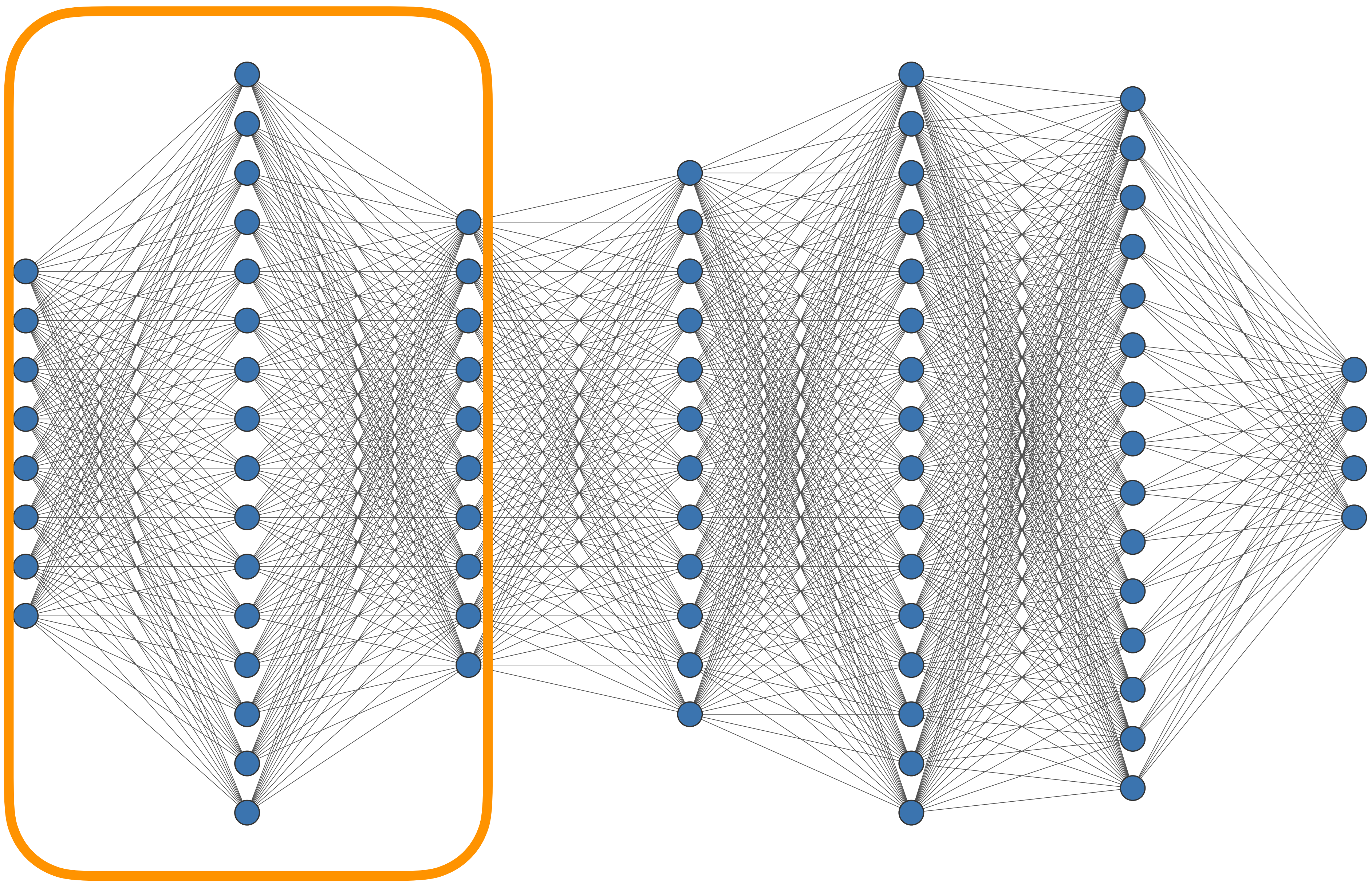


Weight decay favors variation in only a few directions (sparse weights)

Weight decay favors outputs that "share" neurons (sparse neurons)

What Does All of This Mean for Learning With **Deep** Neural Networks?

Layers of Vector-Valued Shallow Networks



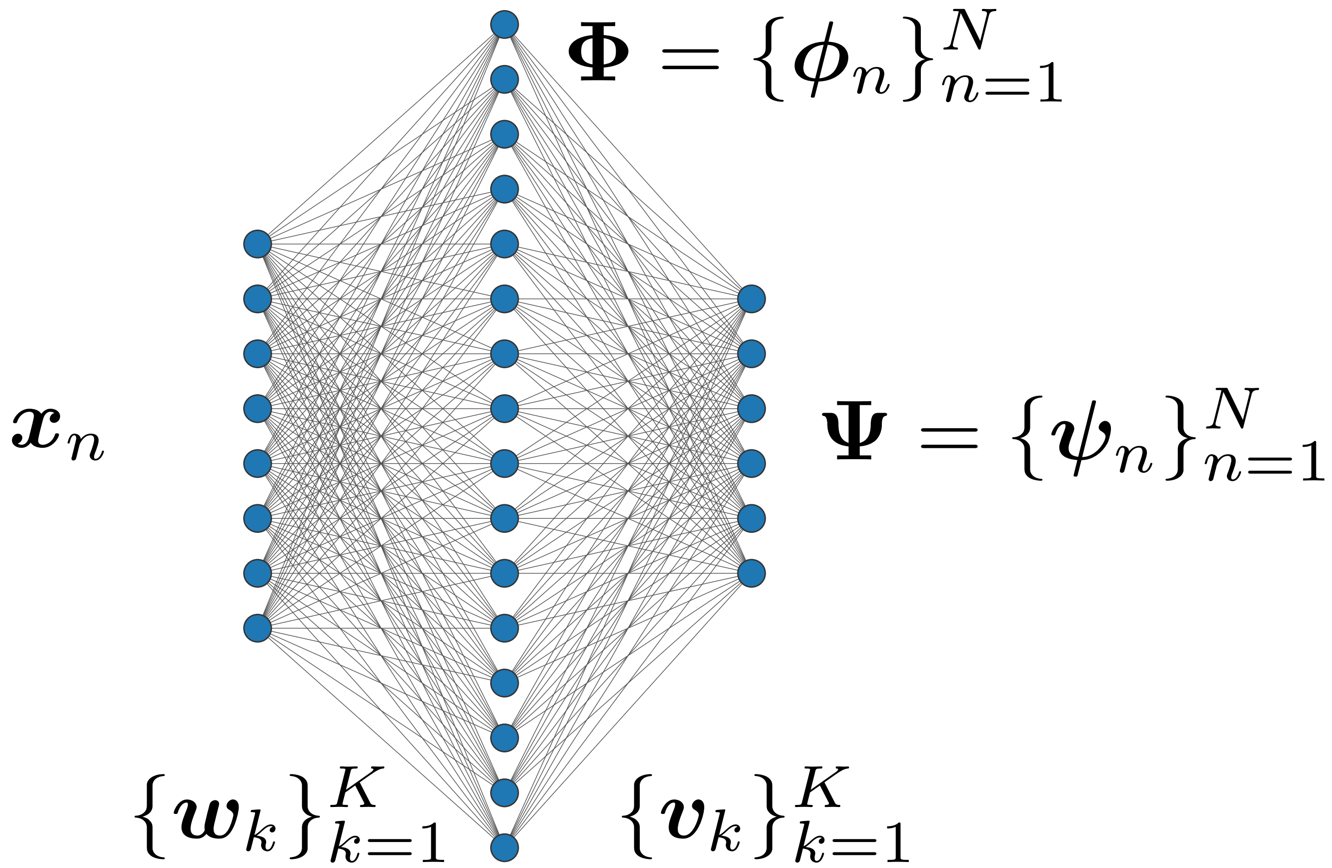
Deep Neural Networks are **Layers** of Shallow Vector-Valued Networks

Tight Bounds on Widths

Consider one ReLU layer within a **trained** deep neural network

with weight decay

to a global minimizer



push the magnitude of $\boldsymbol{w}_k$ into $\boldsymbol{v}_k$

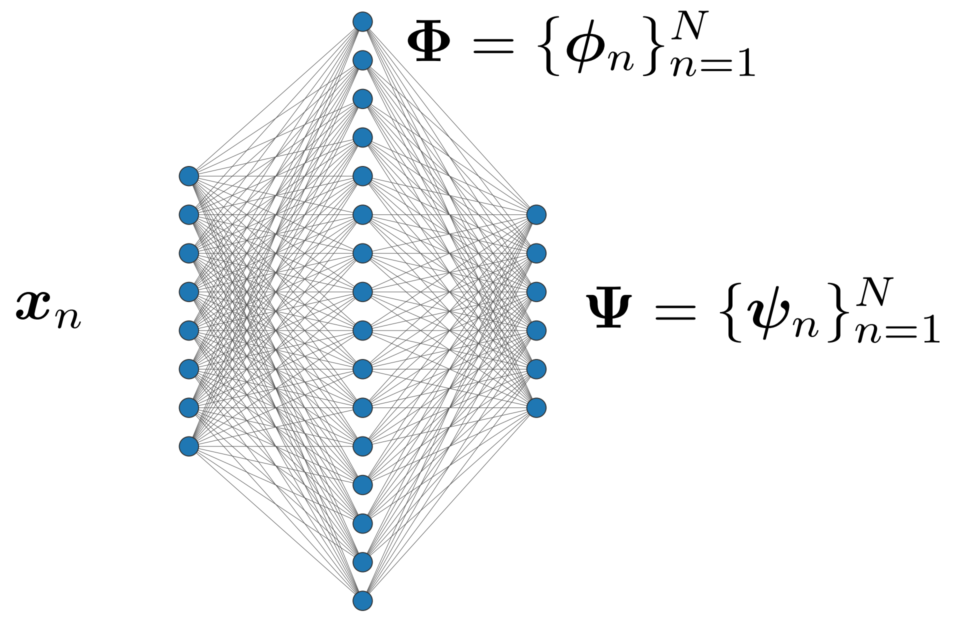
multitask lasso

At each layer, the weight decay solution minimizes

$$\min_{\{\boldsymbol{v}_k\}_{k=1}^K} \sum_{k=1}^K \|\boldsymbol{v}_k\|_2 \quad \text{s.t.} \quad \mathbf{\Psi} = \mathbf{V}\mathbf{\Phi}.$$

Shenouda, **P.**, Lee, and Nowak (2024, JMLR)

Tight Bounds on Widths



$$\min_{\{\boldsymbol{v}_k\}_{k=1}^K} \sum_{k=1}^K \|\boldsymbol{v}_k\|_2 \quad \text{s.t.} \quad \boldsymbol{\Psi} = \mathbf{V}\boldsymbol{\Phi}.$$

Low-rank data embeddings have been observed empirically by Huh et al. (2022).

Layer Width Theorem (Shenouda, **P.**, Lee and Nowak 2024)

Let $\boldsymbol{\Phi}$ denote the post-activation features and $\boldsymbol{\Psi}$ denote the neuron outputs of any ReLU layer in a **trained** DNN (minimizes the weight decay objective). Then, there exists a representation with

$$K \leq \operatorname{rank}(\boldsymbol{\Phi}) \operatorname{rank}(\boldsymbol{\Psi}) \leq N^2$$

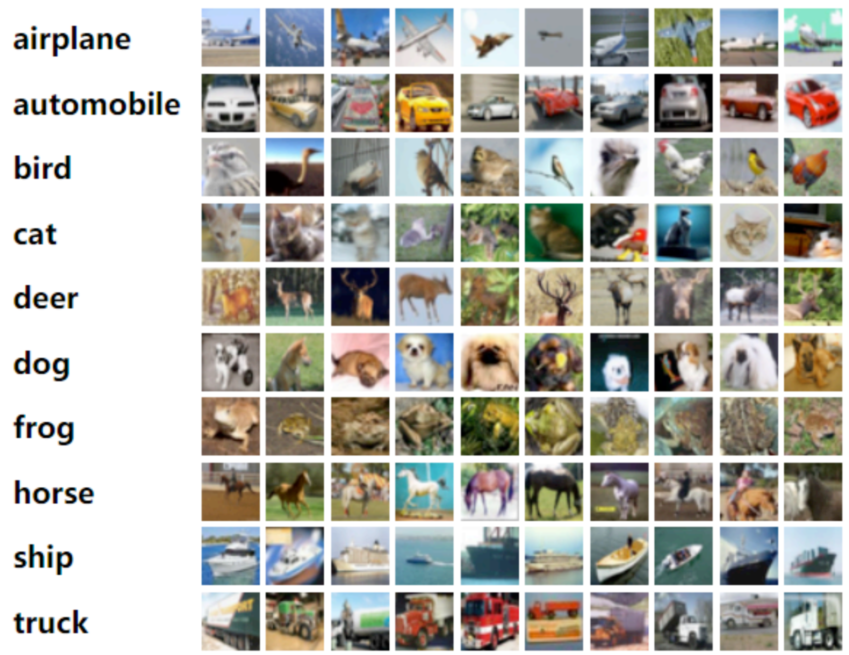
Bound of Jacot et al. (2022): $N(N+1)$.

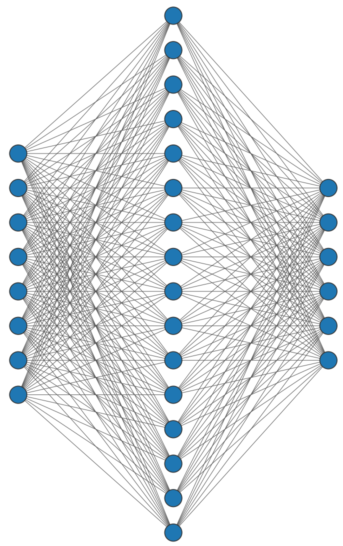
neurons. The representation can be found by solving a **convex multitask lasso** problem.

Shenouda, **P.**, Lee, and Nowak (2024, JMLR)

Application: Principled DNN Compression

VGG-19 trained with weight decay on CIFAR-10.





Theory: There exists a representation with

$$\leq \operatorname{rank}(\boldsymbol{\Phi}) \operatorname{rank}(\boldsymbol{\Psi}) \approx 10 \cdot 10 = 100 \text{ neurons.}$$

	original network	compressed network
active neurons	512	47
test accuracy	93.92%	93.88%
train loss	0.0104	0.0112

$10\times$ compression!

no change in performance!

Shenouda, **P.**, Lee, and Nowak (2024, JMLR)