Coding Schemes in Non-Lazy Artificial Neural Networks

Alexander van Meegen[1,2] & Haim Sompolinsky[1,3]

1) Center for Brain Science, Harvard; 2) Brain Mind Institute, EPFL; 3) ELSC, Hebrew University.

Question

What is the structure of learned representations in artificial neural networks?

Approach

Consider fully connected feedforward networks in the non-lazy regime which enforces strong representation learning.

Non-lazy: readout scales its inputs with 1/N *after* learning the task, $f(x)=\frac{1}{N}\sum_{i=1}^{N}a_i\phi[z_i(x)]$

Weights drawn from Bayes posterior

$$P(\Theta)=\frac{1}{Z}\exp\left[-\beta\mathcal{L}(\Theta)+\log P_0(\Theta)\right]$$

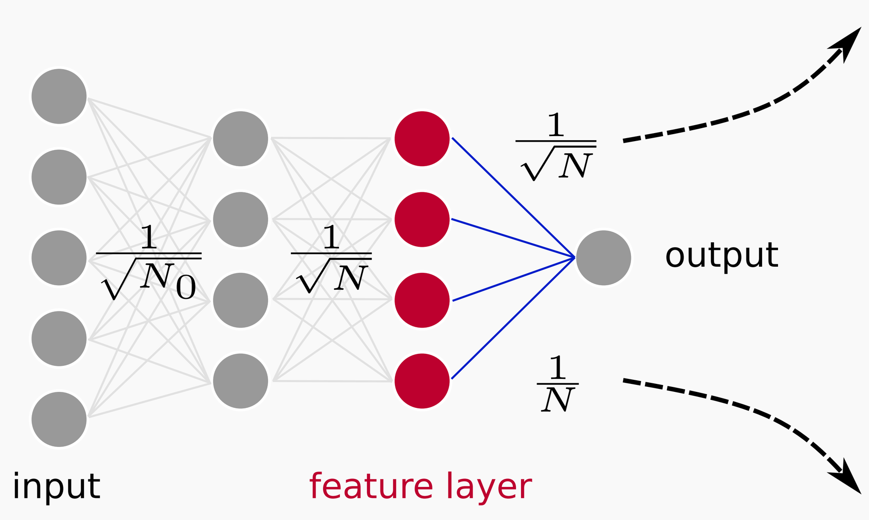
loss Gaussian prior

Focus: zero temperature limit (β→∞) which enforces zero (MSE) loss

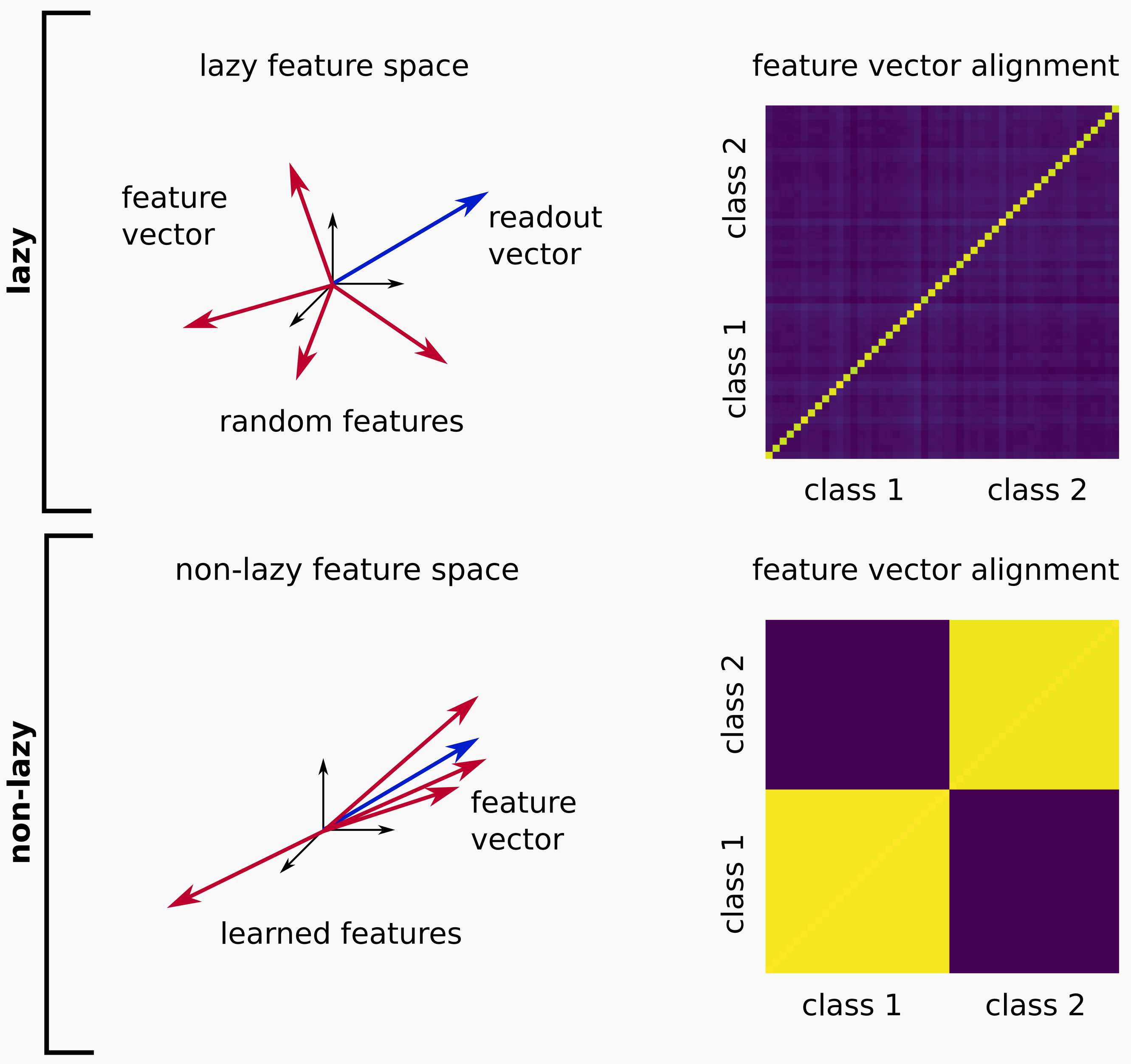
Theory: number of neurons N, training set size P, and input dimensionality to infinity at fixed ratio

Overview of lazy and non-lazy regimes

"Laziness" controlled by the scaling of the output at initialization [1-3] ...



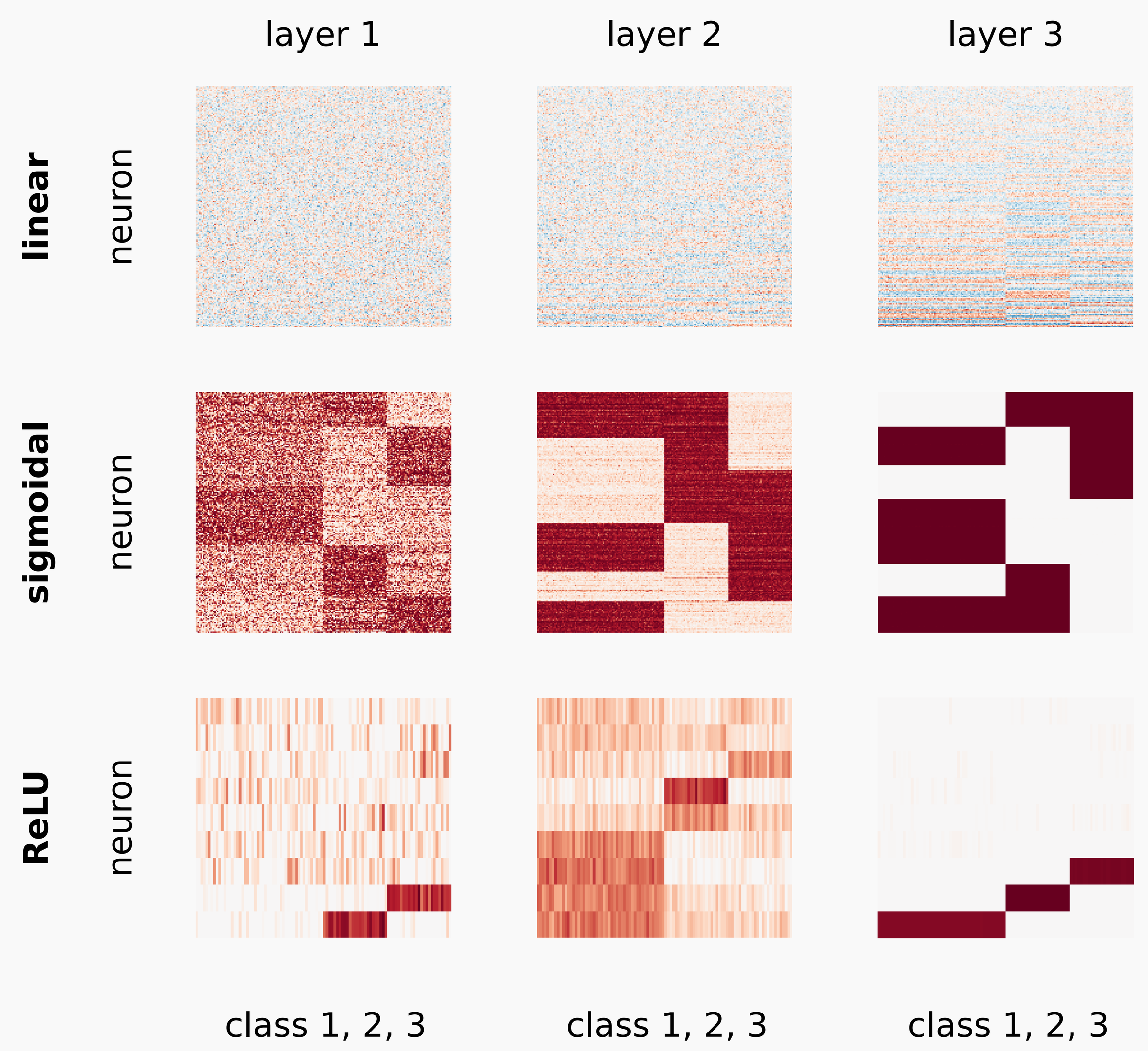
... and after learning, i.e., drawn from the weight posterior.



Coding schemes (classification)

Here: 3 classes with unequal ratio, orthogonal data.

Salient structure in neural activations, depends strongly on neuronal non-linearity:



Analog coding scheme (linear):
all neurons respond to all classes but with different strengths.

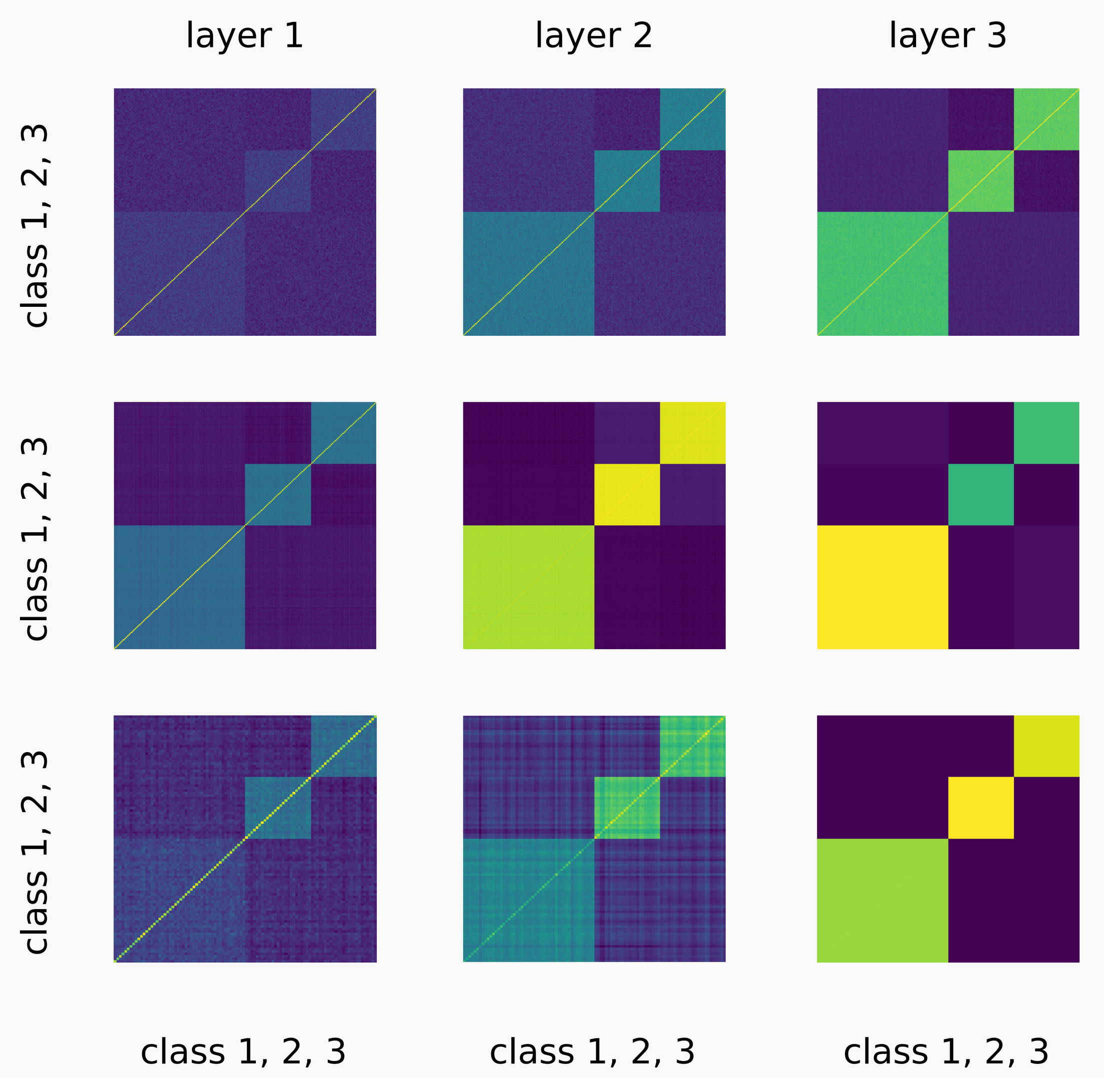
Redundant coding scheme (sigmoidal):
all the codes in the scheme are shared by a large subset of neurons.

many neurons with code 1-1-0

Sparse coding scheme (ReLU)
only a small subset of neurons exhibit codes.

single neuron with code 1-0-0

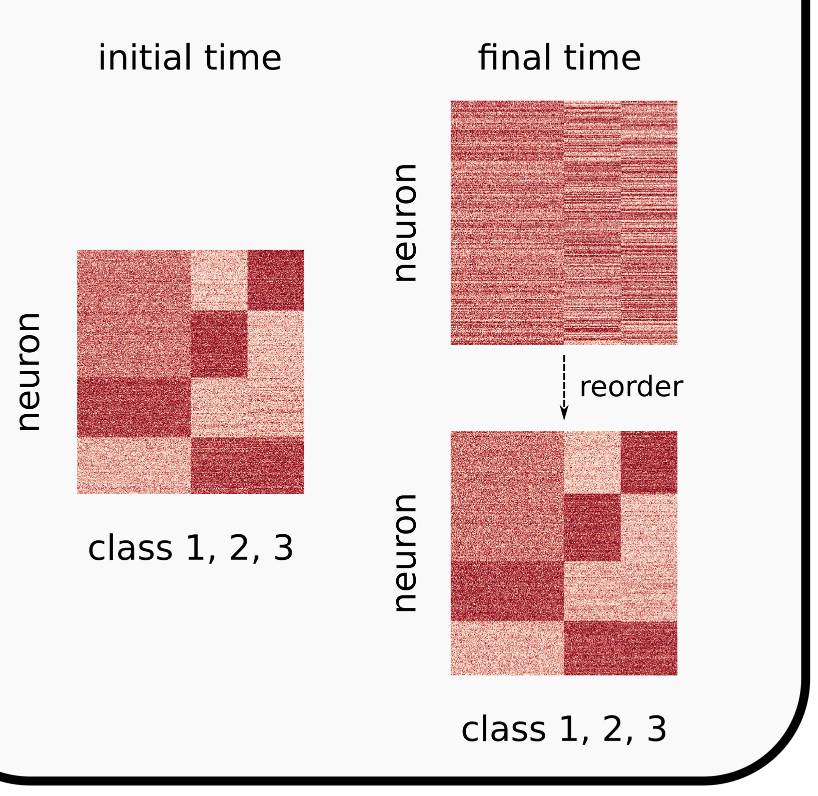
Geometry of representations (kernel) reflects task structure, similar for all neuronal non-linearities:



Drifting representations

Origin of drift: sampling dynamics.

Linear: drifting representations in all layers.
ReLU: no drift (symmetry breaking).
Sigmoidal: no drift in feature layer (symmetry breaking), drift in lower layers:

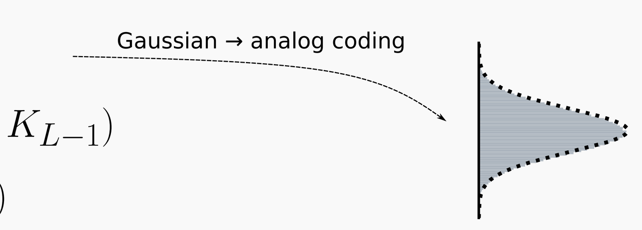


Theory (executive summary)

Main result: posterior of readout weights and preactivations factorizes across layers and neurons into single-neuron posteriors.

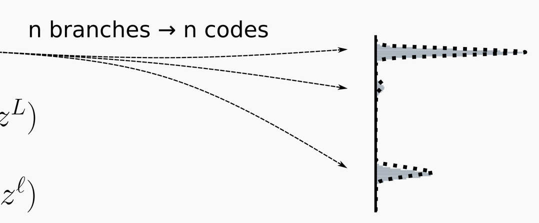
linear

$$P(a)=\mathcal{N}(a\,|\,0,U)$$
$$P(z^L\,|\,a)=\mathcal{N}(z^L\,|\,YU^{-1}a,K_{L-1})$$
$$P(z^\ell)=\mathcal{N}(z^\ell\,|\,0,K_\ell)$$



sigmoidal

$$P(a)=\sum_{\gamma=1}^{n}P_\gamma\delta(a-a_\gamma)$$
$$P(z^L\,|\,a)\propto\mathcal{N}(z^L\,|\,0,K_{L-1})e^{a^\top t^\top\phi(z^L)}$$
$$P(z^\ell)\propto\mathcal{N}(z^\ell\,|\,0,K_{\ell-1})e^{\frac{1}{2}\phi(z^\ell)^\top\tilde{K}_\ell\phi(z^\ell)}$$



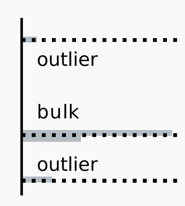
ReLU

bulk

$$P(a)=\delta(a-a_0)$$
$$P(z\,|\,a)\propto\mathcal{N}(z\,|\,0,K_0)e^{a^\top t^\top\phi(z)}$$

outlier

$$P(a_i)=\delta(a_i-\sqrt{N}\bar{a}_i)$$
$$P(z_i\,|\,a_i)=\delta(z_i-\sqrt{N}\bar{z}_i)$$



Important detail: scaling of prior readout weight variance with data set size P.

Not shown: parameters of single-neuron posteriors detemined self-consistently.

Take home

Summary: theory of learned representations deep in the feature learning regime.

Representations are embedded into *coding schemes* in a classification context, the details depend on the neuronal non-linearity.

Emergence of high-level *category-selectivity* in the feature layer of a simple neural network.

Permutation-symmetry breaking controls the presence or absence of drift during sampling.

Not shown: *generalization* beyond training examples and corresponding representations; MNIST and CIFAR10 examples.

References: [1] Chizat, Oyallon, Bach; NeurIPS (2019). [2] Mei, Montanari, Nguyen; PNAS (2018). [3] Bordelon & Pehlevan; NeurIPS (2022).

Acknowledgments: This work has been supported by the Swartz Foundation, the Gatsby Charitable Foundation, and the Kempner Institute.