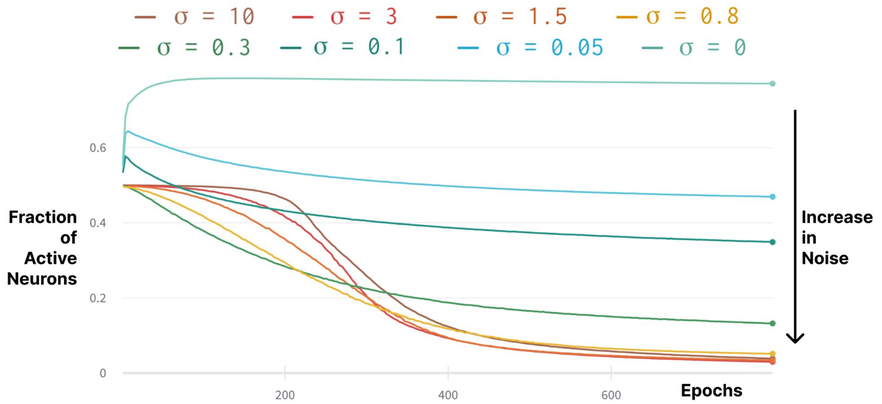
# Emergence of Sparse Representations from Noise

Trenton Bricken[1,2], Rylan Schaeffer[3], Bruno Olshausen[2], Gabriel Kreiman[4]

[1] Systems, Synthetic and Quantitative Biology, Harvard University;
[2] Redwood Center for Theoretical Neuroscience, University of California, Berkeley;
[3] Computer Science, Stanford University;
[4] Programs in Biophysics and Neuroscience, Harvard Medical School.

## 1. Noise ➔ Sparsity:

**Fraction of Active Neurons as a function of Gaussian Noise during training**

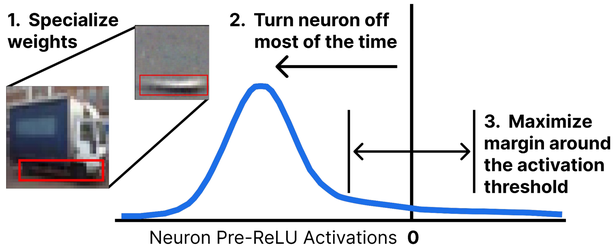



Key Findings:

- Works for reconstruction and classification tasks
- Removing noise post-learning results in the network remaining sparse
- Applies across shallow and deep MLPs including Transformers (but not AlexNet)
- And datasets:
  - MNIST, CIFAR10, WikiText-103
- And activation functions:
  - GeLU, Elu, Tanh, Sigmoid (ask about the latter two)

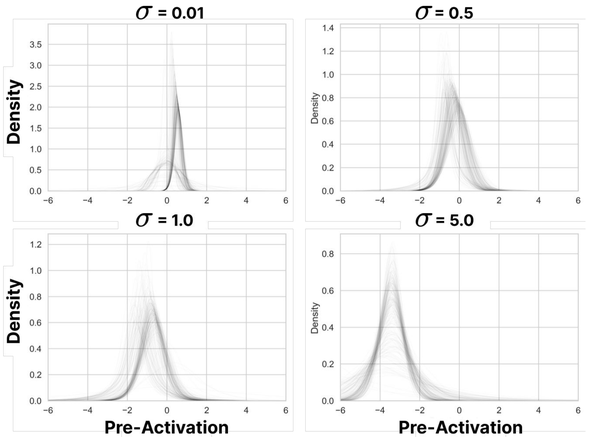
## 2. Noise ➔ Three Implicit Loss Terms:

Theoretical Distribution of Neuron Activations

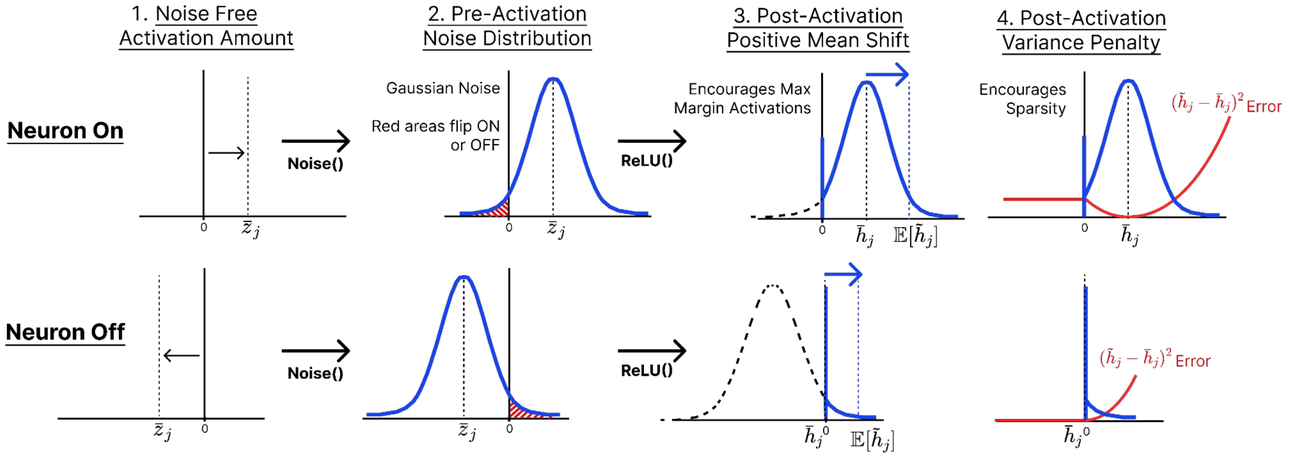



Noise Terms ➔ Maximize(
[1. Neuron Sparsity] +
[2. Neuron Activation Margin] +
[3. Specialize Model Weights])

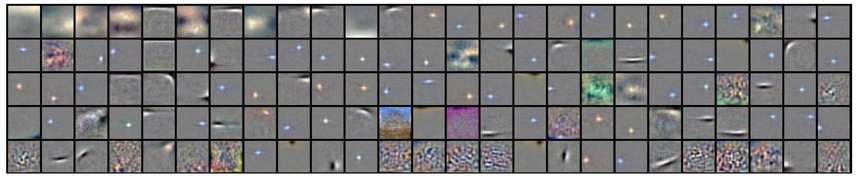
Empirical Neuron Activations by Noise Scale




## 3. Noise ➔ Flips Neurons On/Off:




## 4. Noise ➔ Biological Receptive Fields



Receptive fields for the neurons most activated by a CIFAR10 truck image include both Gabor filters (present in V1) and on/off center/surround (present in LGN)

Key Findings:

- All bias terms and weight norms synchronize to a single scalar
  - Replacing biases with a single shared bias scalar has min. effect
- This results in a natural Top-K activation function emerging, similar to an inhibitory interneuron
- Top-K and Sparse coding (L1 penalty) don't produce on/off center/surround, only Gabors

***Why are brains sparse?***

Canonical Answer: Metabolic efficiency – Action Potentials are expensive!
Complimentary Answer: Great for de-noising – biological systems and sensory data are very noisy!

**Paper**