Neural networks trained with SGD learn distributions of increasing complexity

Maria Refinetti (ENS Paris & G-Research)
Alessandro Ingrosso (ICTP Trieste)
Sebastian Goldt (SISSA)

ICML '23

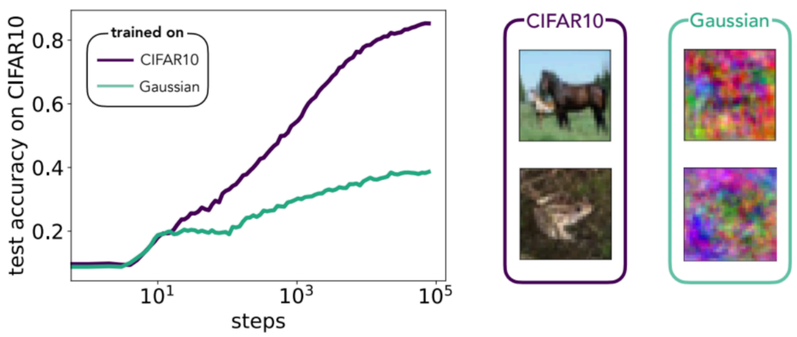
In a Pitaya

The Gaussian world is not enough! Are you really surprised?

Maybe.. because **Gaussian equivalence principle :** test error of a neural network trained on realistic inputs can be exactly captured by a Gaussian model of the data

→ And I trust these guys!

DenseNet121 test accuracy on CIFAR10



Higher **O**rder **C**umulants (HOC) improve generalisation (even in two-layer NN!)

But... we do not know why! :(

Our Question: How do NN learn about HOCs?

Results : Distribution Simplicity Bias (DSP)

i.e. mean and covariance

Neural Networks trained using SGD first classify their inputs using **lower-order input** statistics. They exploit **higher-order statistics** only **later during training**

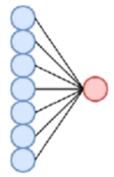
Higher order cumulants

- demonstrate DSB in a solvable model of a single neurone trained on synthetic data
- demonstrate DSB empirically in various deep convolutional networks and visual transformers trained on CIFAR10
- demonstrate that it even holds in networks pre-trained on ImageNet

Theoretical work : A day in the life of a perceptron

Setup

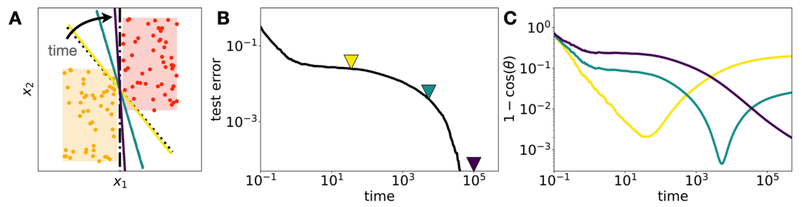
- Perceptron
- Online learning: limite of infinite data



$$\hat{y} = \sigma(\lambda) \qquad \lambda \equiv w_i x^i / \sqrt{D}$$

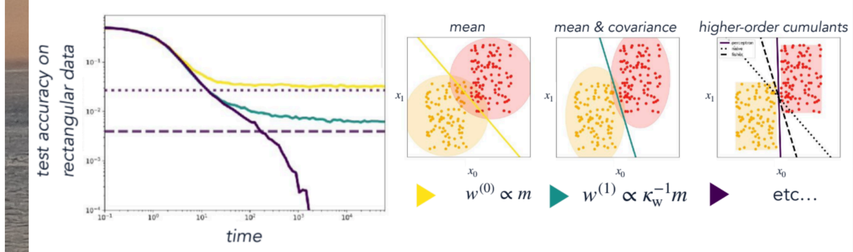
Gradient Flow $\tau \dot{w}_i = -\mathbb{E}\,\nabla_{w_i} L(w)$

A perceptron learns to take increasingly higher-order statistics into account



- **0th order** in w $\quad w_i^{(0)} \propto m^i \equiv \kappa_+^i - \kappa_-^i$
- **1st order** $\quad w_i^{(1)} \propto (\kappa_w)_{ij} m^j$
- **3th order** $\quad w_i^{(c)} \propto -(\kappa_w)_{ij} \kappa_w^{j,k,l,m} w_k^{(1)} w_l^{(1)} w_m^{(1)}$

A perceptron *sees* distributions of increasing complexity

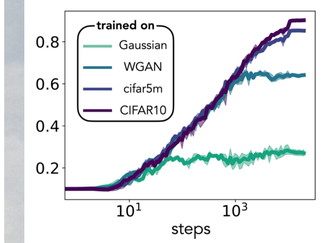


Empirical Results

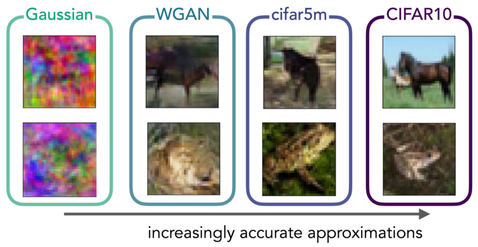
NN learn distributions of increasing complexity

A simplicity bias in neural networks

Resnet18 test accuracy on CIFAR10

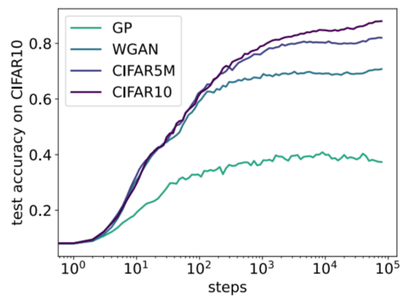


Training distributions (CIFAR10 and the "clones")



Pre-trained NN learn DIC, but faster

Pre-training on ImageNet does not change pattern in Resnet18



→ Even a Resnet18 pre-trained on ImageNet learns distributions of increasing complexity (but faster)

Conclusion

How do Neural Networks learn about Higher order cumulants?

- Neural Netoworks learn distributions of increasing complexity
- Even a perceptron will go beyond Gaussian equivalent models if they are task-relevant!
- GF analysis tells us which statistics are key.