Rainbow Networks: A Model of Learned Weights in Deep Networks

Florentin Guth[1,2] Brice Ménard[3] Gaspar Rochette[4] Stéphane Mallat[2, 5]

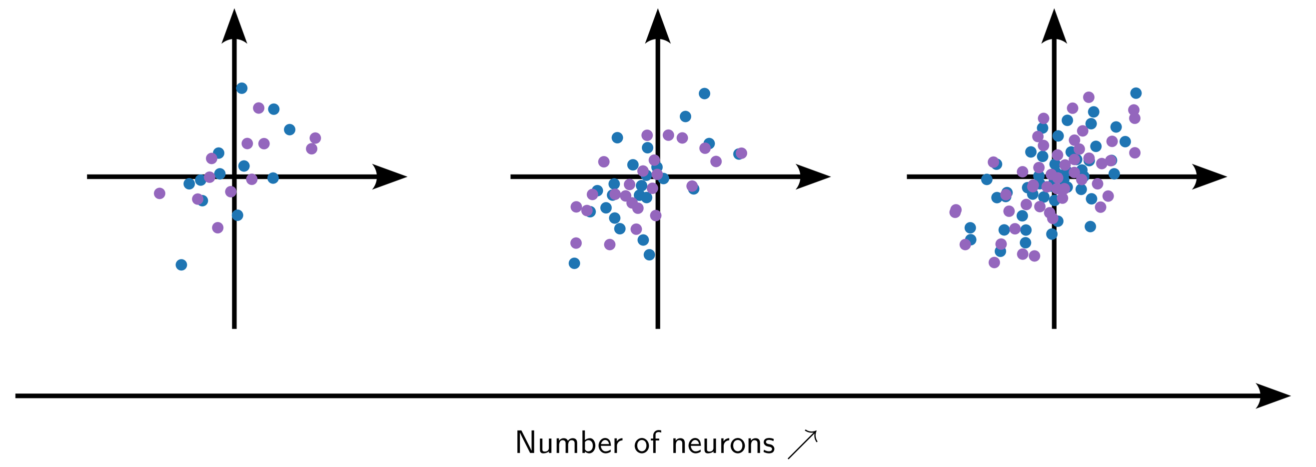
[1]New York University, New York, USA [2]Flatiron Institute, New York, USA [3]Johns Hopkins University, Baltimore, USA [4]DI, ENS, CNRS, PSL University, Paris, France [5]Collège de France, Paris, France

Introduction & Summary

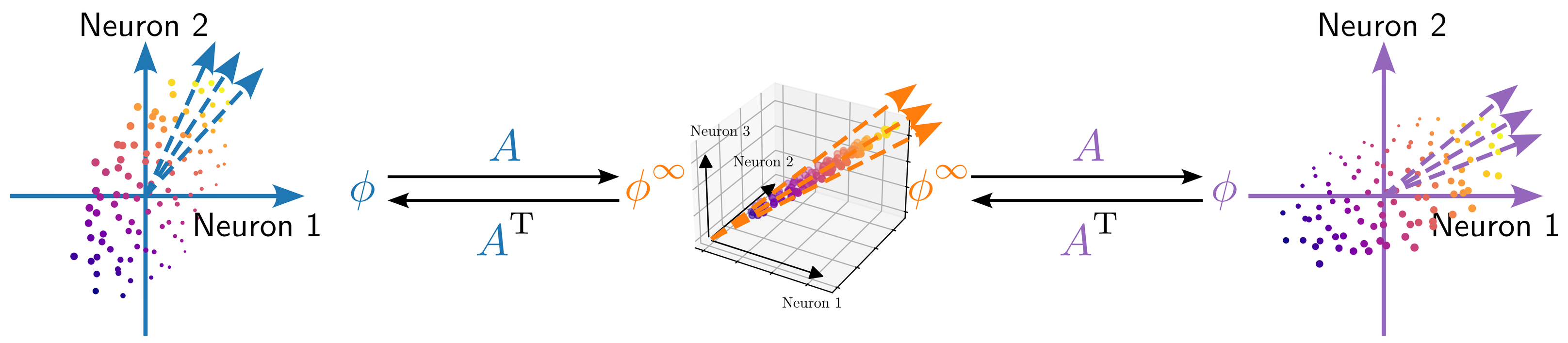
- Deep networks are considered black boxes. Mathematical models based on random weights, but either kernel regime (NNGP, NTK) or one-hidden-layer (mean-field)
- How to understand feature learning in deep networks? **What is the probability distribution of trained weights?**
- We introduce the **rainbow model** of deep networks: random features but with dependencies between layers
- Rainbow networks can achieve **comparable accuracies as trained networks**

Key Concept: Aligned Weight Distributions

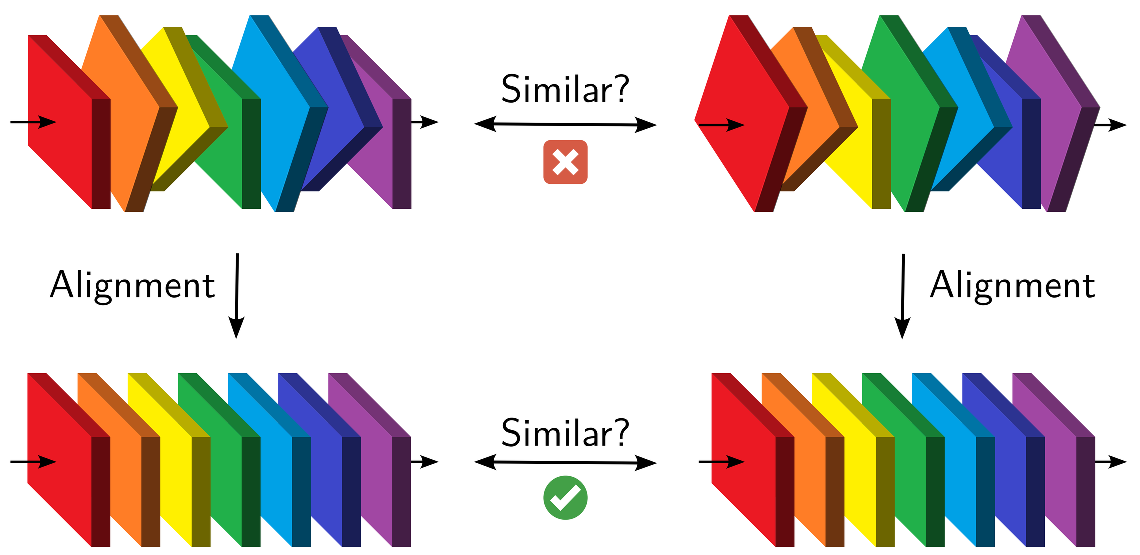
Mean-field picture: neurons as samples (random features)



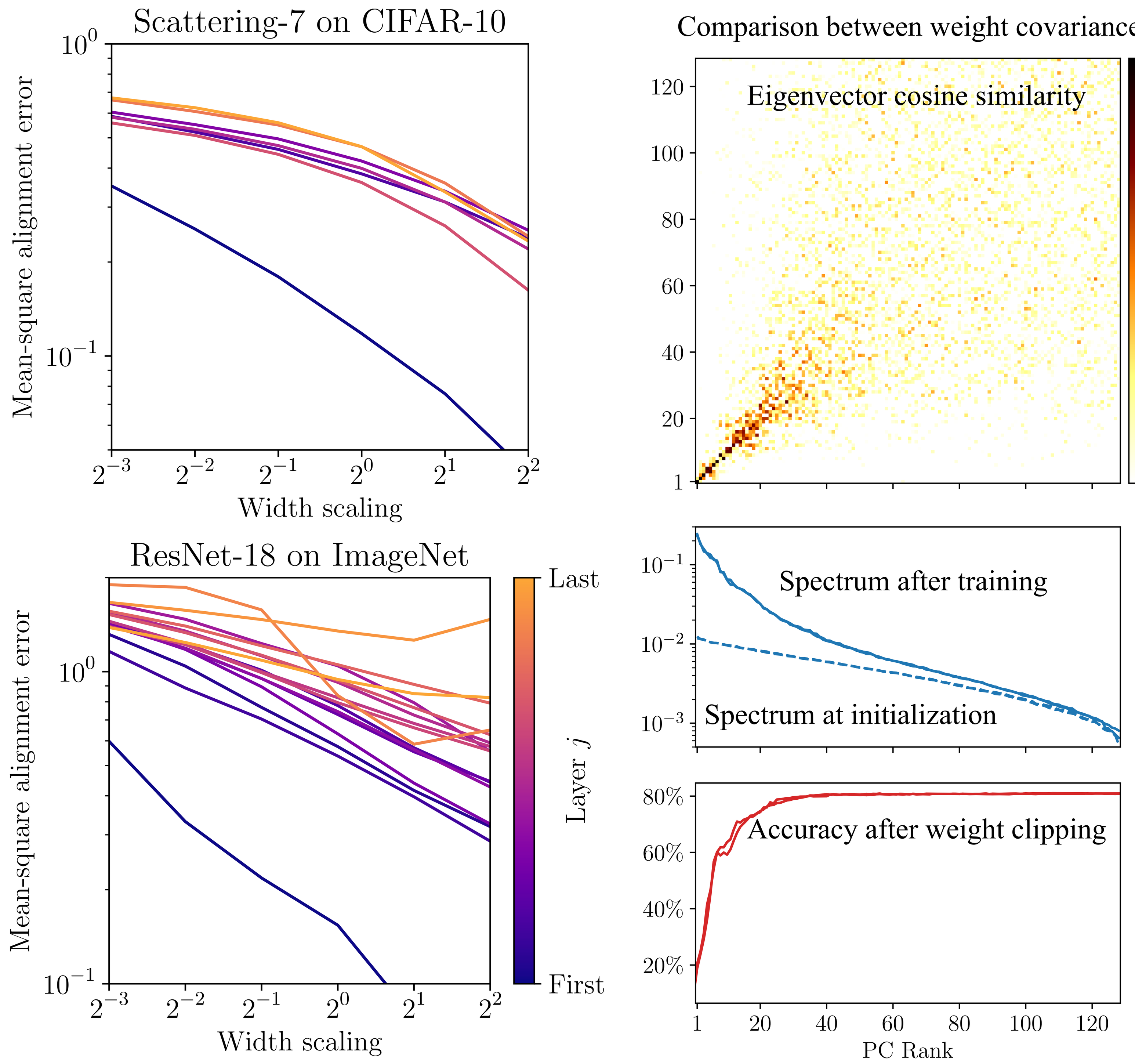
Alignment: hidden representations are approximately equal up to a rotation



Summary: **aligned** networks have similar activations and weights



CNNs Learn the Same Weights After Alignment



Alignment Convergence of Rainbow Networks

- A random feature map $\phi(x) = \left(n^{-1/2}\sigma(\langle x, w_i\rangle)\right)_{i\leq n}$ with i.i.d. $w_i \sim \pi$ defines a kernel $\langle\phi(x),\phi(x')\rangle = \frac{1}{n}\sum_{i=1}^n \sigma(\langle x, w_i\rangle)\,\sigma(\langle x', w_i\rangle)$
- The law of large number implies it converges towards

$$\mathbb{E}_{w\sim\pi}[\sigma(\langle x, w\rangle)\,\sigma(\langle x', w\rangle)] = \langle\phi^\infty(x),\phi^\infty(x')\rangle$$

- For large widths, **activations are fixed up to a rotation**:

$$\langle\phi(x),\phi(x')\rangle \approx \langle\phi^\infty(x),\phi^\infty(x')\rangle \implies \begin{cases} A\,\phi(x) \approx \phi^\infty(x) \\ \phi(x) \approx A^{\mathrm{T}}\phi^\infty(x)\end{cases}$$

- **Theorem:** *there exists a closed-form orthogonal A such that*

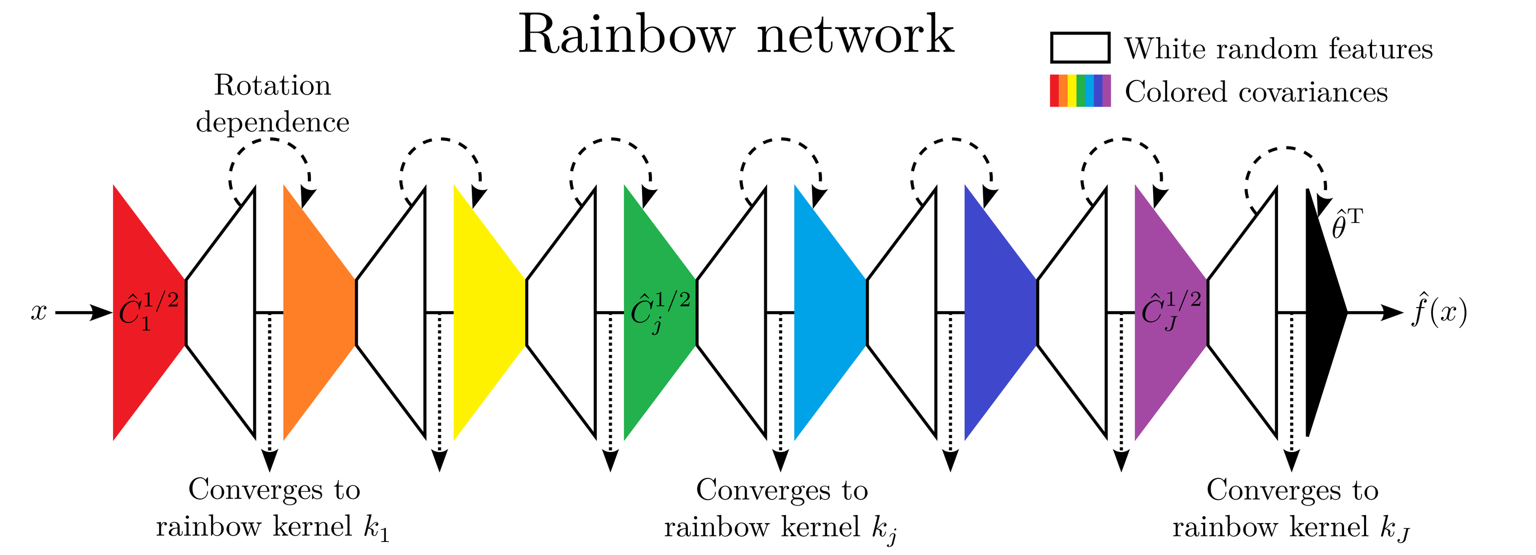
$$\mathbb{E}_{W,x}\left[\|A\,\phi(x) - \phi^\infty(x)\|^2\right] = O(n^{-\gamma})$$

Assumptions: π has finite fourth-order moments + capacity condition.

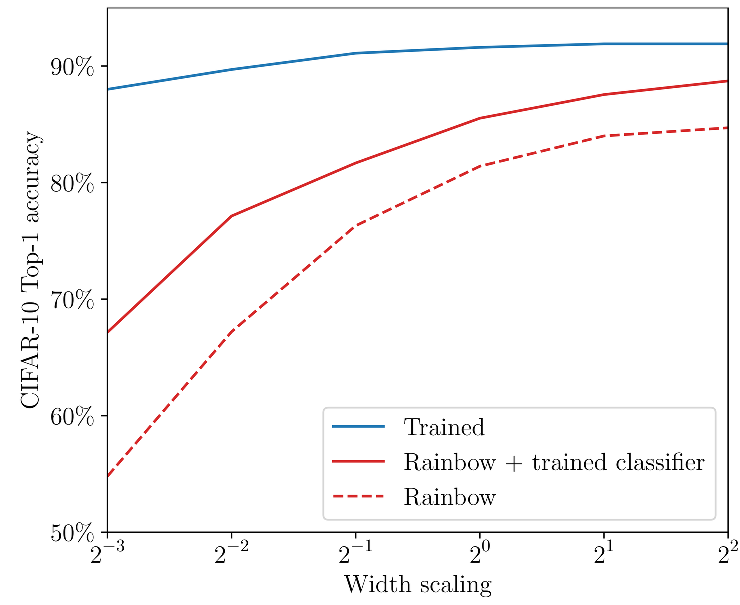
- **Next layer weights can cancel the rotation**: if $w = A^{\mathrm{T}} w'$, then

$$\langle\phi(x), w\rangle = \left\langle\phi(x), A^{\mathrm{T}} w'\right\rangle = \langle A\,\phi(x), w'\rangle \approx \langle\phi^\infty(x), w'\rangle$$

Gaussian Rainbow Networks

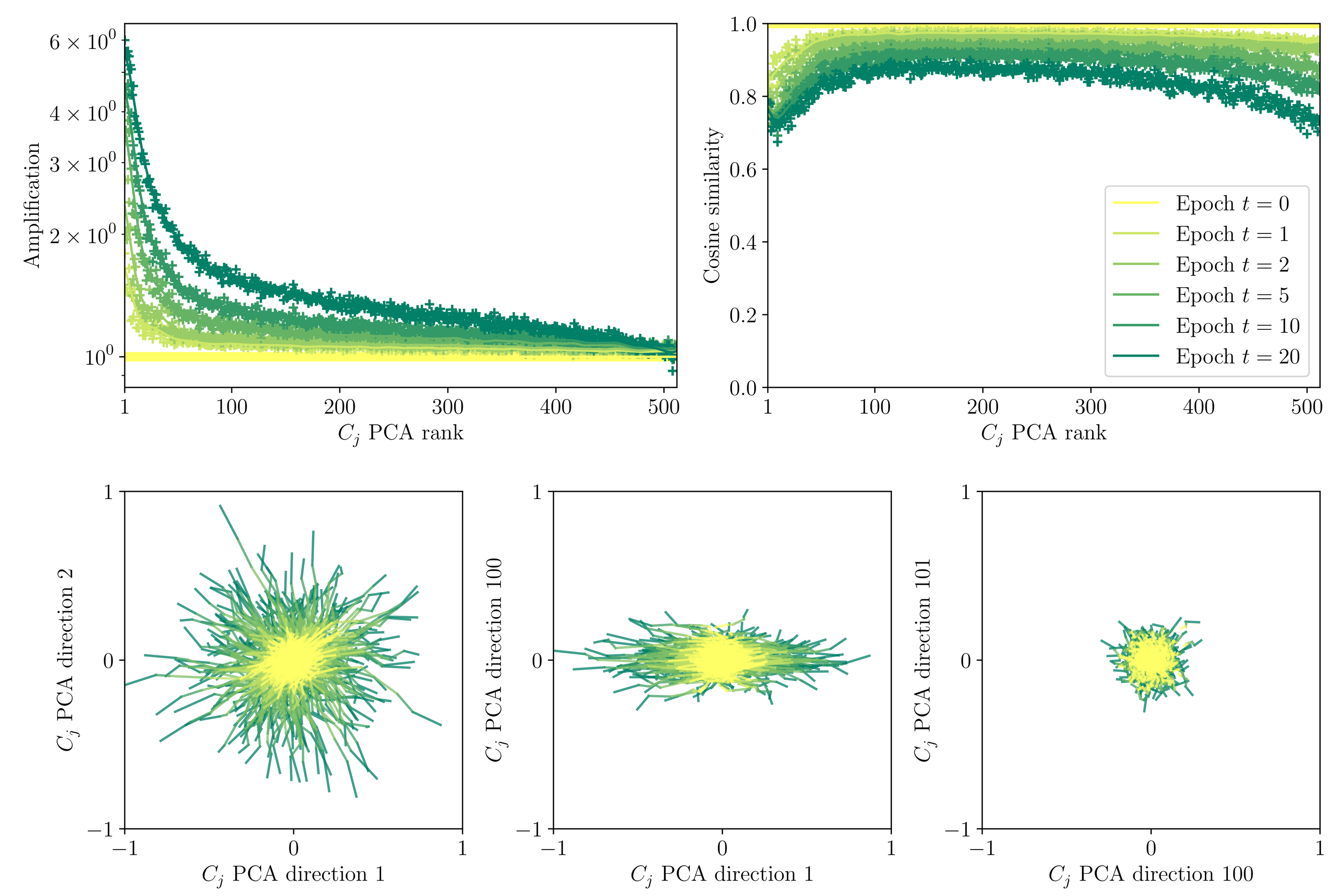


- The Gaussian rainbow model is defined from a trained reference network with activations $\phi_\ell^\infty(x)$ and weight covariances C_ℓ at each layer ℓ
- Sample new weights $w_{\ell,i}$ from $\mathcal{N}(0, C_\ell)$ and align them at each layer
- Test on CNNs with learned channels weights but fixed spatial filters



Insights Into Training Dynamics

During training, weights are linearly stretched without internal motion:



florentin.guth@nyu.edu

https://bonnerlab.github.io/ccn-tutorial/pages/analyzing_neural_networks.html

https://arxiv.org/abs/2305.18512