Data-driven emergence of convolutional structure in neural networks

A. Ingrosso, S. Goldt

PNAS 119 (40), 2022

Data symmetries and neural networks

Exploiting invariances in the data is key for efficient learning

Symmetries
in the data

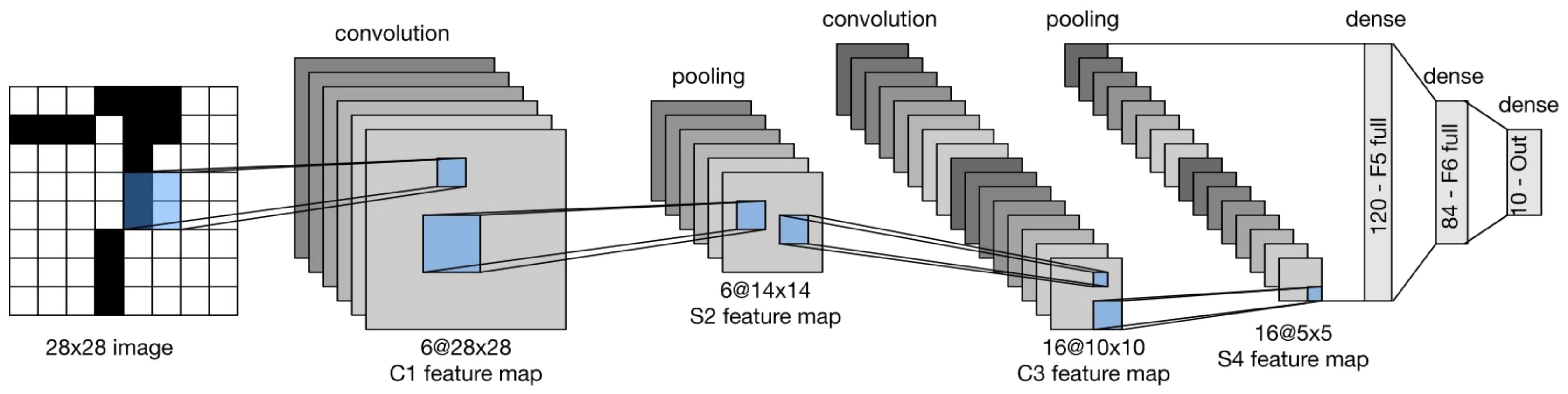
Data symmetries and neural networks

Exploiting invariances in the data is key for efficient learning

Symmetries
in the data

+

Appropriate network
architecture



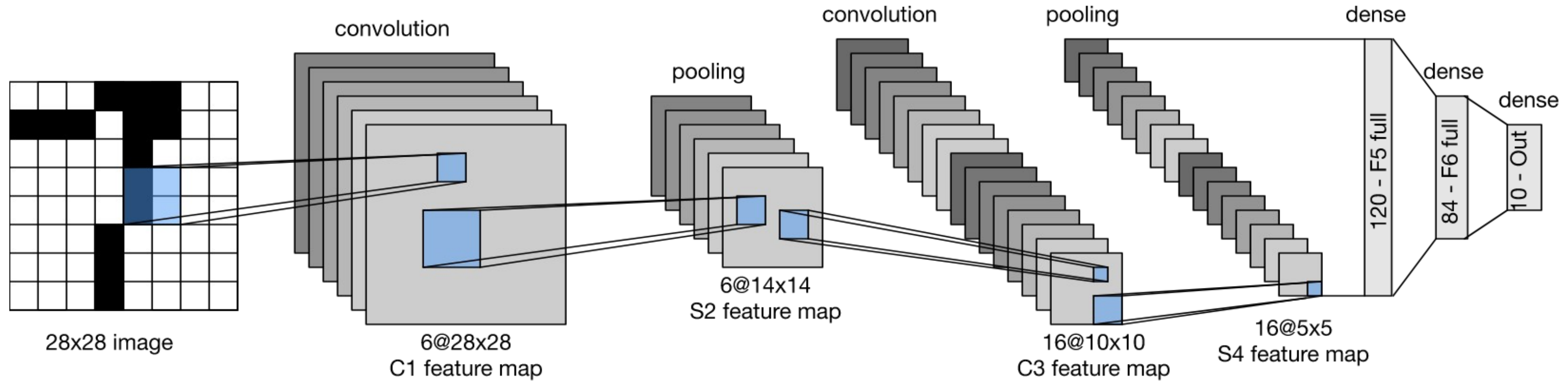
Data symmetries and neural networks

Exploiting invariances in the data is key for efficient learning

Symmetries
in the data

+

Appropriate network
architecture



↓

Sample/parameter-efficient
learning

ConvNets and Fully Connected networks

- **Hallmarks of convolutions**
  - Local connectivity with weight sharing
  - Tessellation of input space

- **Fully-connected networks**
  perform worse on image classifications tasks

Can we learn a convolutional structure from scratch?

A minimal model of natural images

High-dimensional dataset with tunable higher-order spatial correlations

Translation-invariant Gaussians:

label index

$$\left\langle z_i^\mu \right\rangle = 0$$

$$\left\langle z_i^\mu z_j^\mu \right\rangle = e^{-(|i-j|/\xi^\mu)^2}$$

correlation length

pass Gaussians through nonlinearity

$$\boldsymbol{x}^\mu = \frac{\psi(g\boldsymbol{z}^\mu)}{Z(g)}$$

A minimal model of natural images

High-dimensional dataset with tunable higher-order spatial correlations

label index

Translation-invariant Gaussians:

$$\langle z_i^\mu \rangle = 0$$

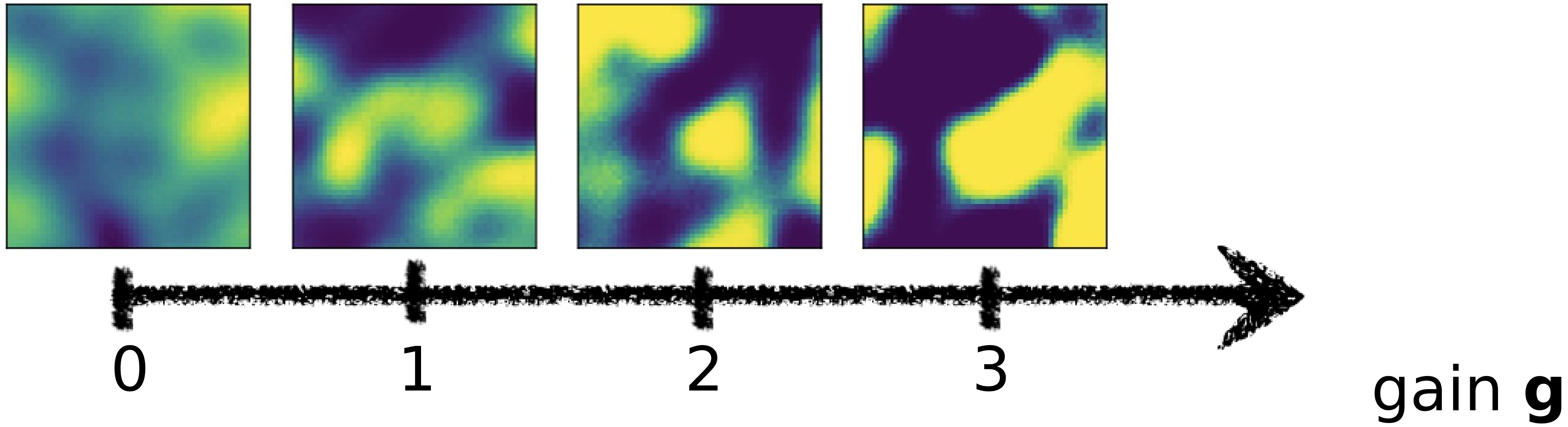
$$\left\langle z_i^\mu z_j^\mu \right\rangle = e^{-(|i-j|/\xi^\mu)^2}$$

correlation length

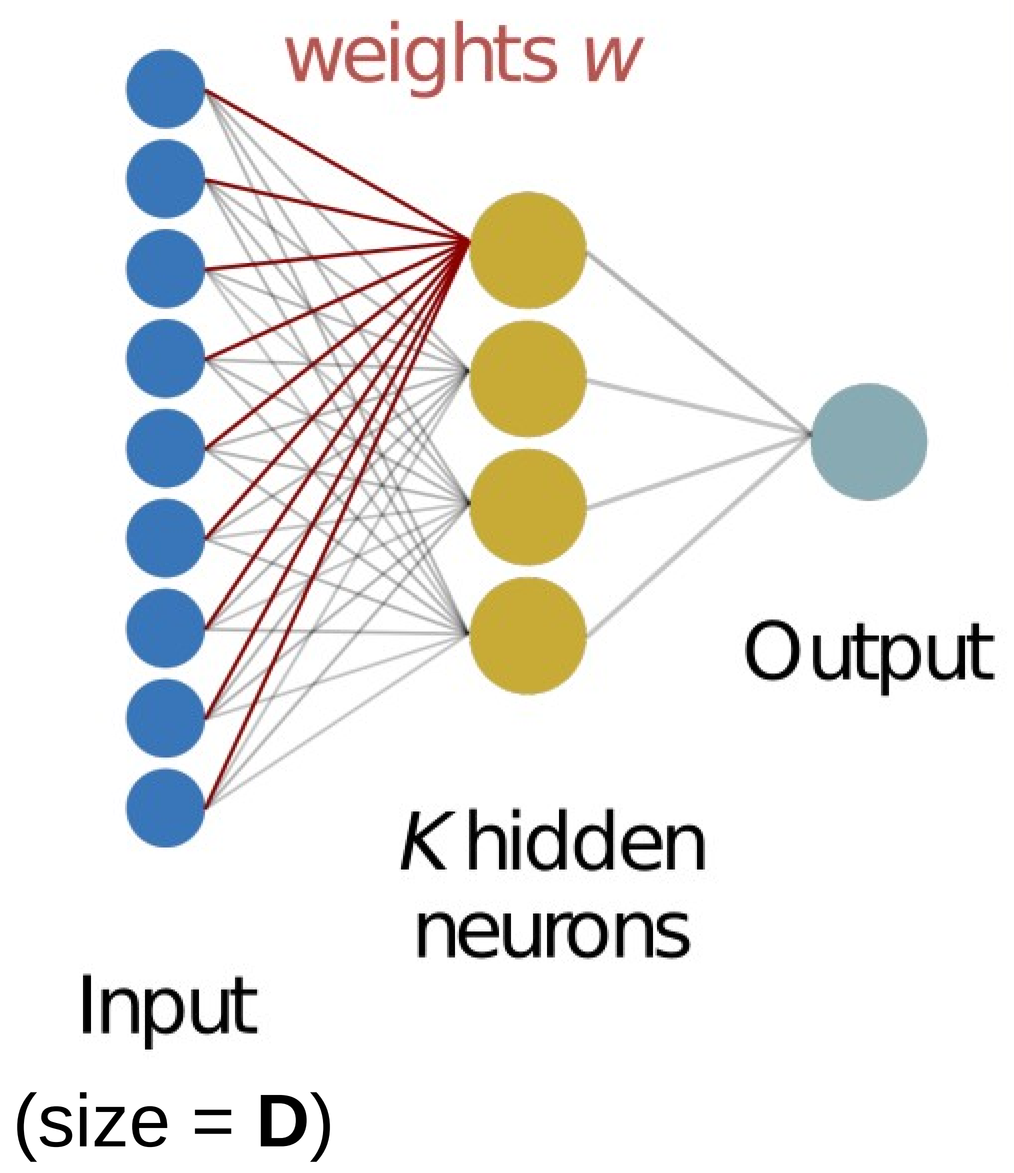
pass Gaussians through nonlinearity →

$$\boldsymbol{x}^\mu = \frac{\psi(g\boldsymbol{z}^\mu)}{Z(g)}$$

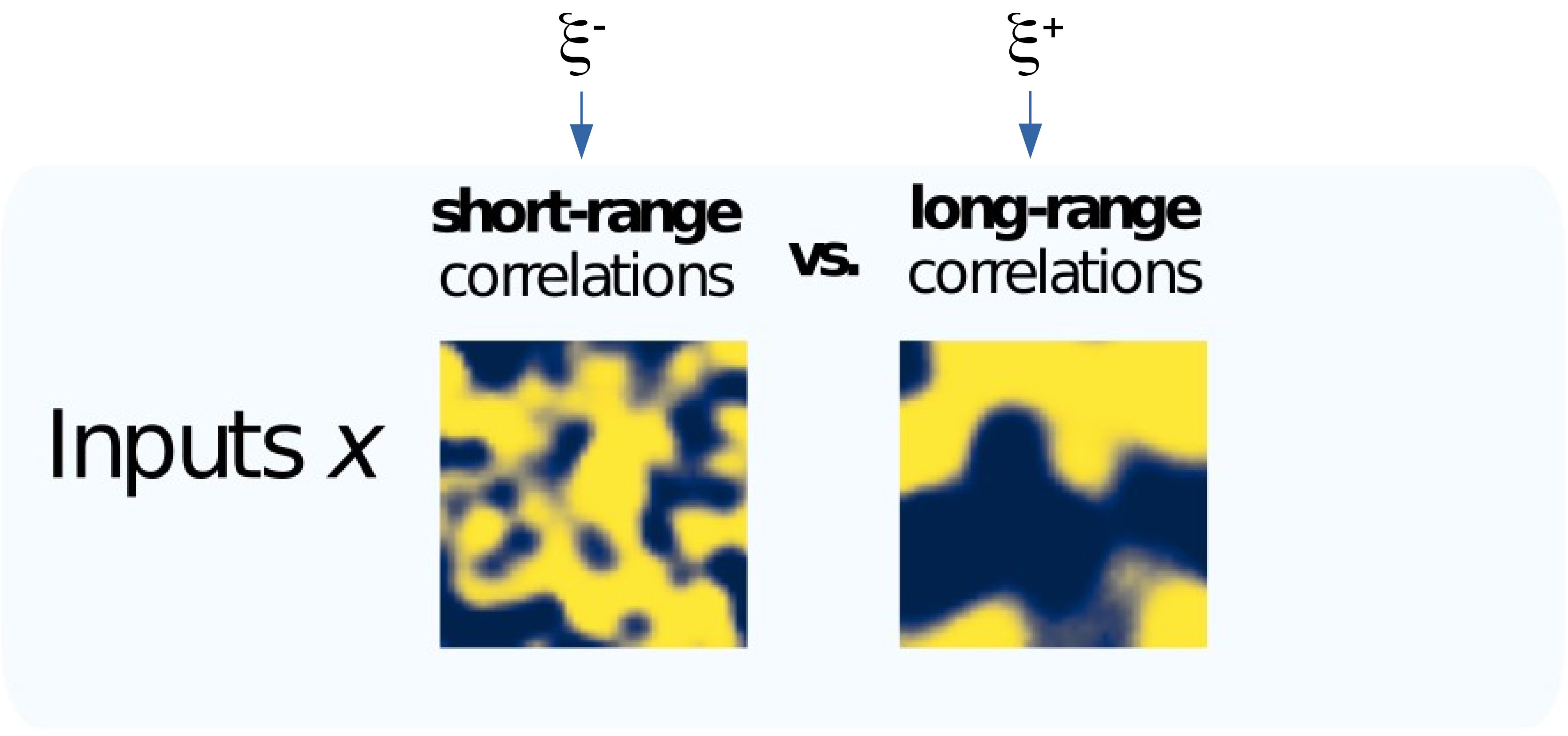
Sample images:



Discriminating “images” with different correlation lengths

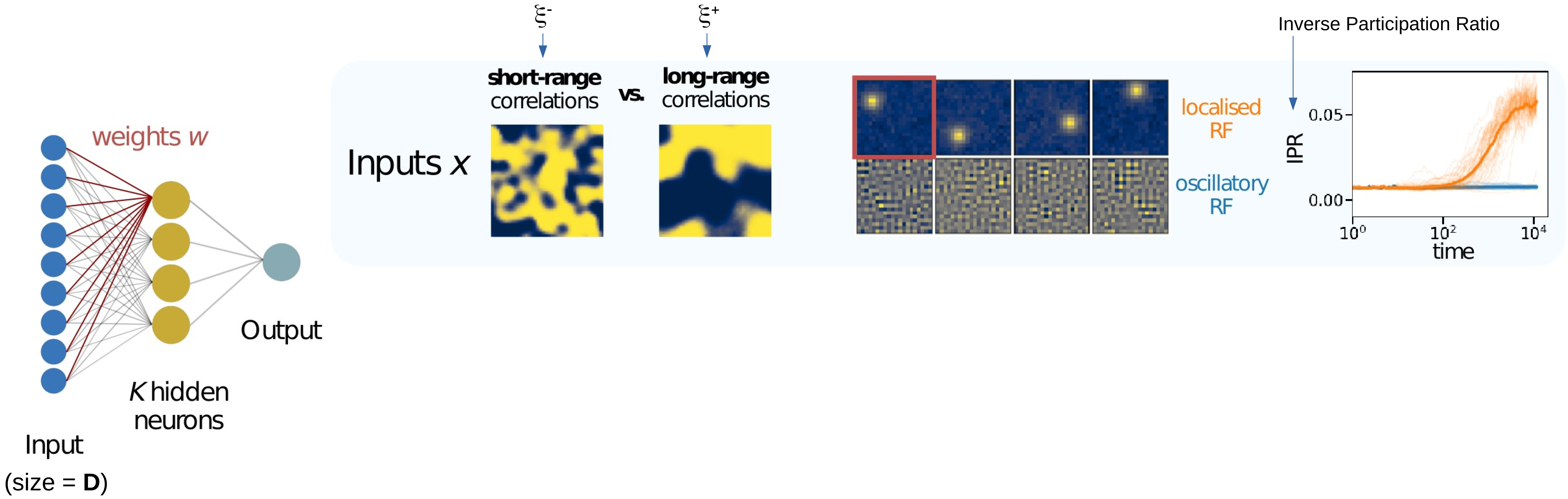


Network architecture



Training inputs

Discriminating “images” with different correlation lengths



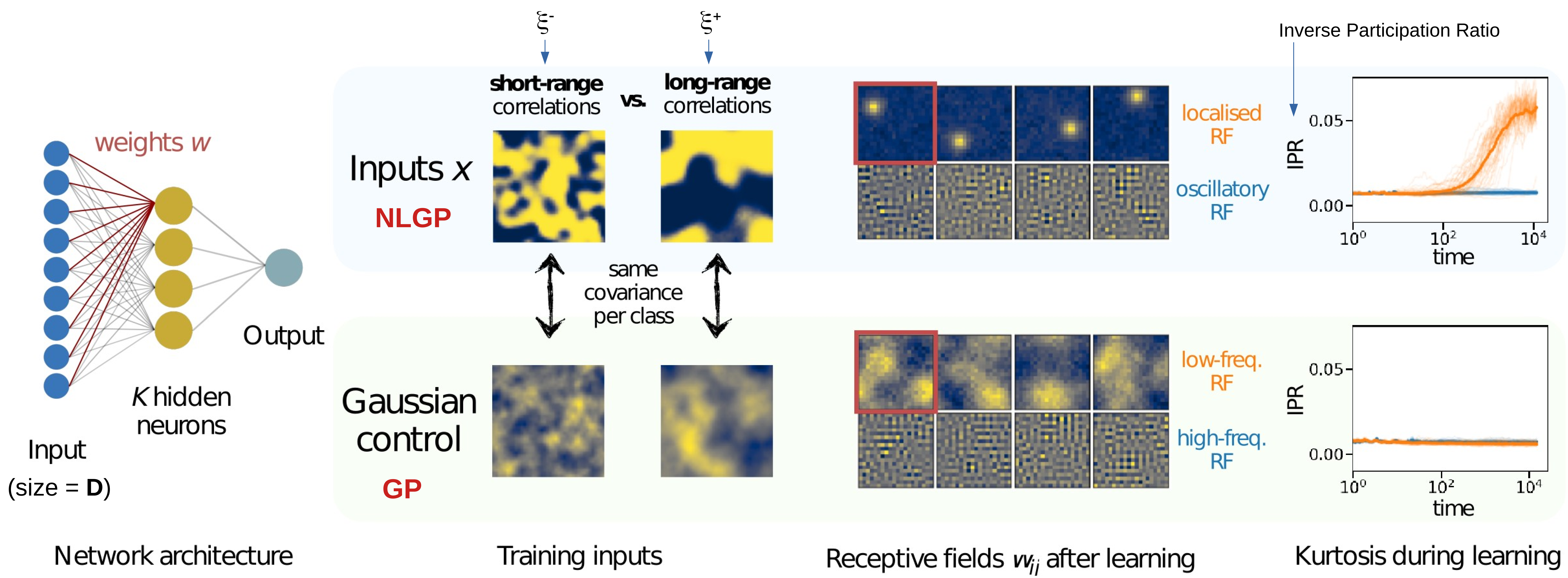
Network architecture

Training inputs

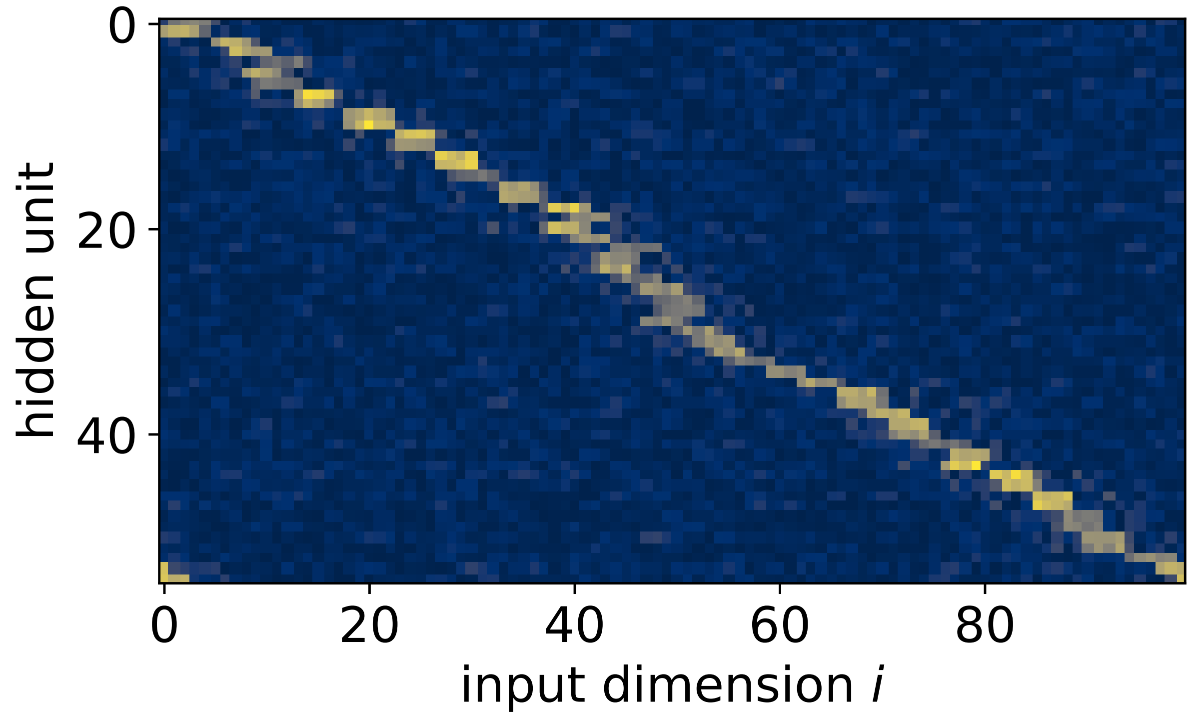
Receptive fields w_{ij} after learning

Kurtosis during learning

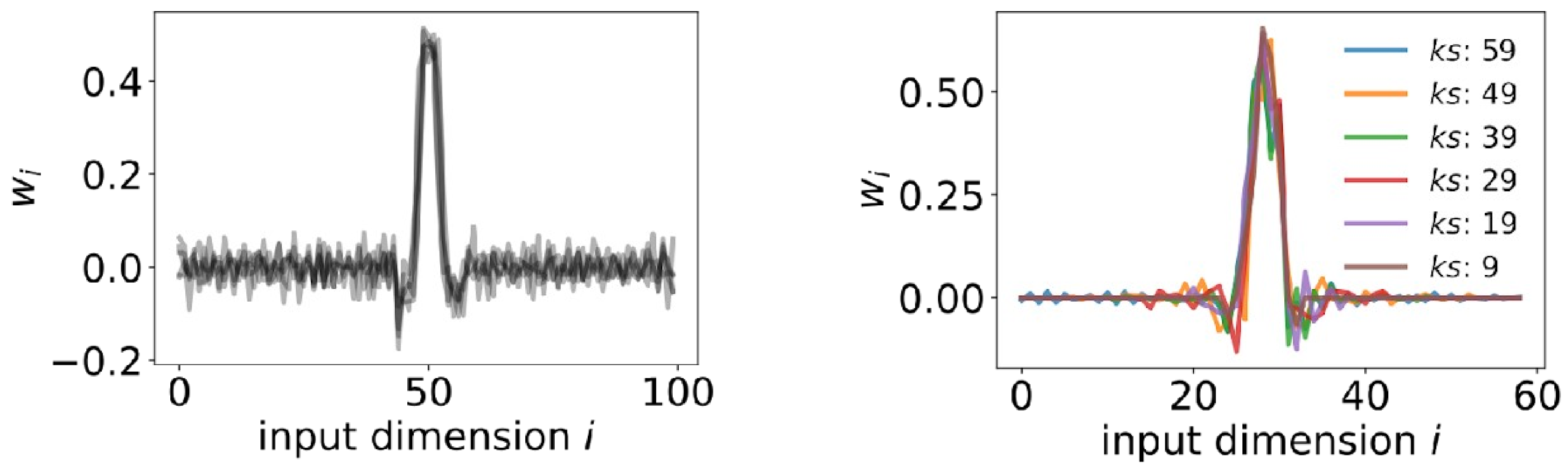
Discriminating “images” with different correlation lengths



Receptive fields tile input space and resemble convolutional filters



Localised receptive fields tesselate input space

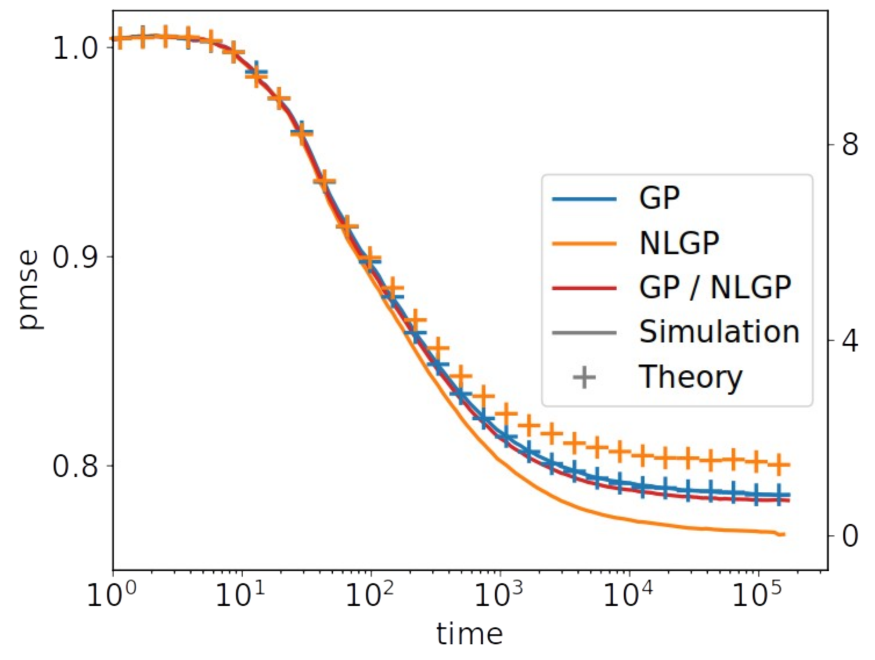


Learnt weight vectors resemble filters of convolutional networks

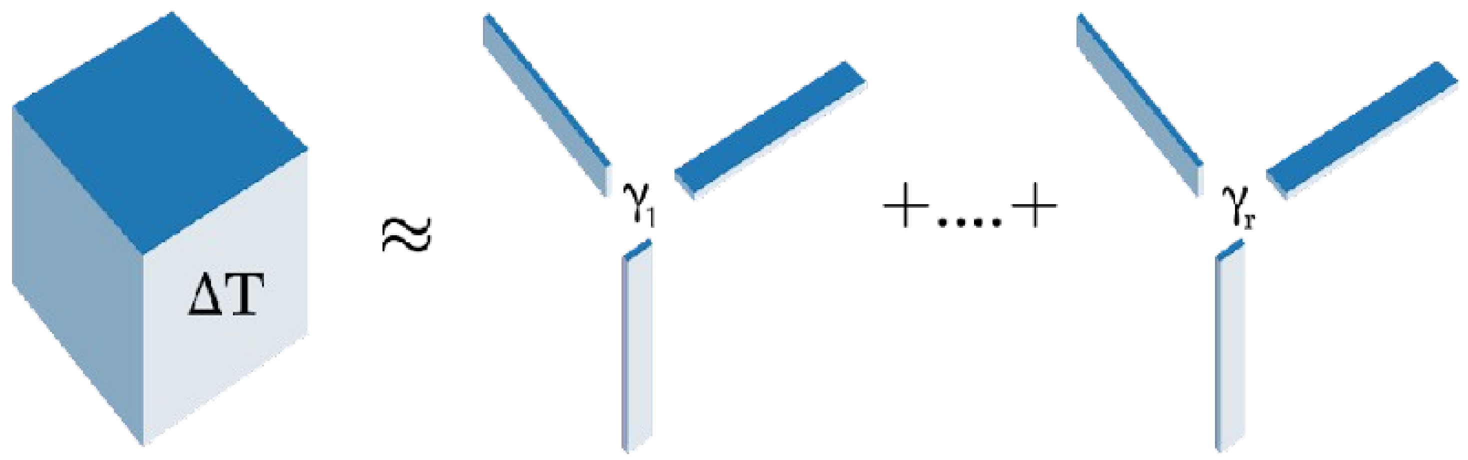
WHAT’S GOING ON

Long story short

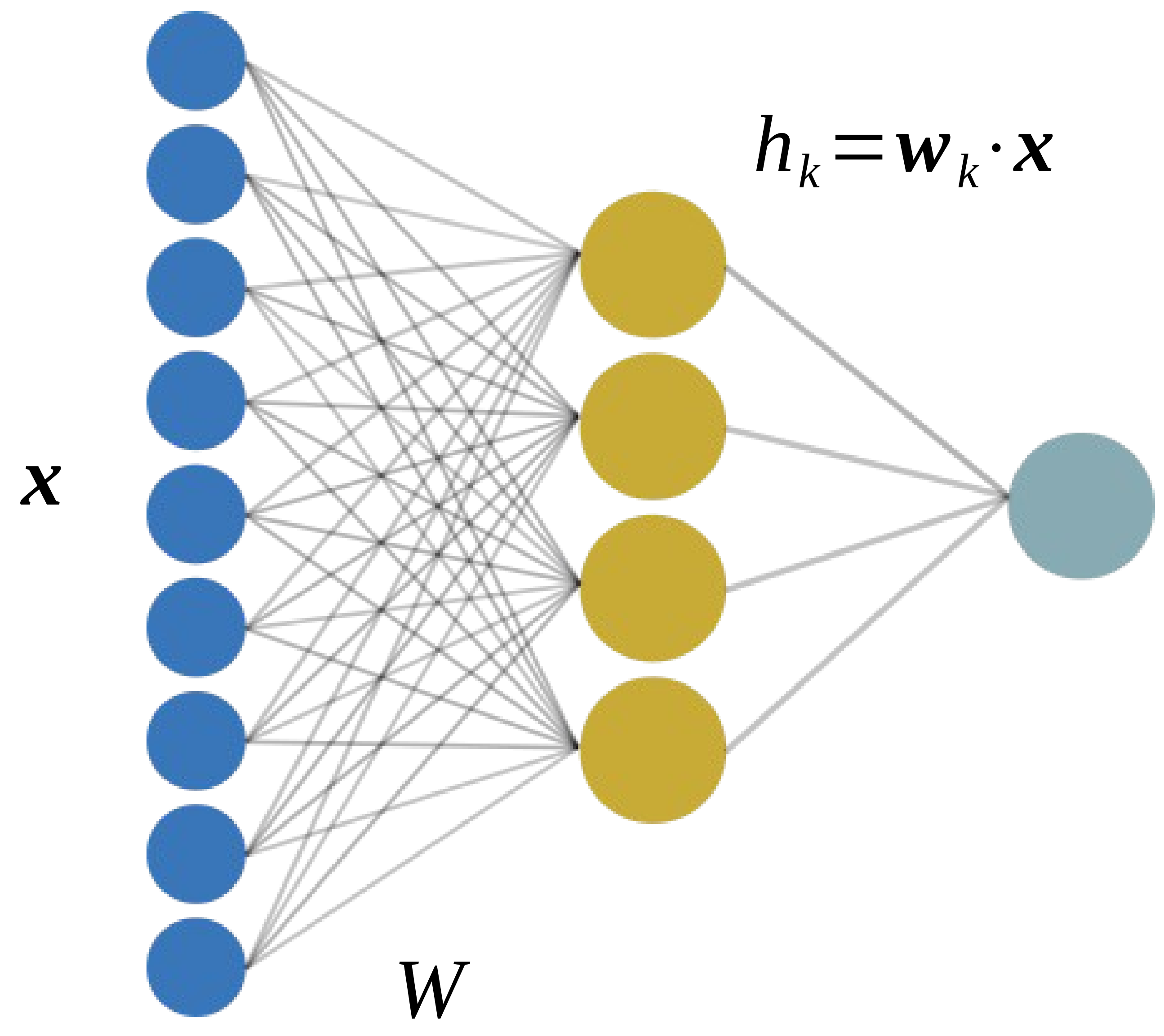
Standard Gaussian theories fail in predicting learning dynamics and formation of localized RF



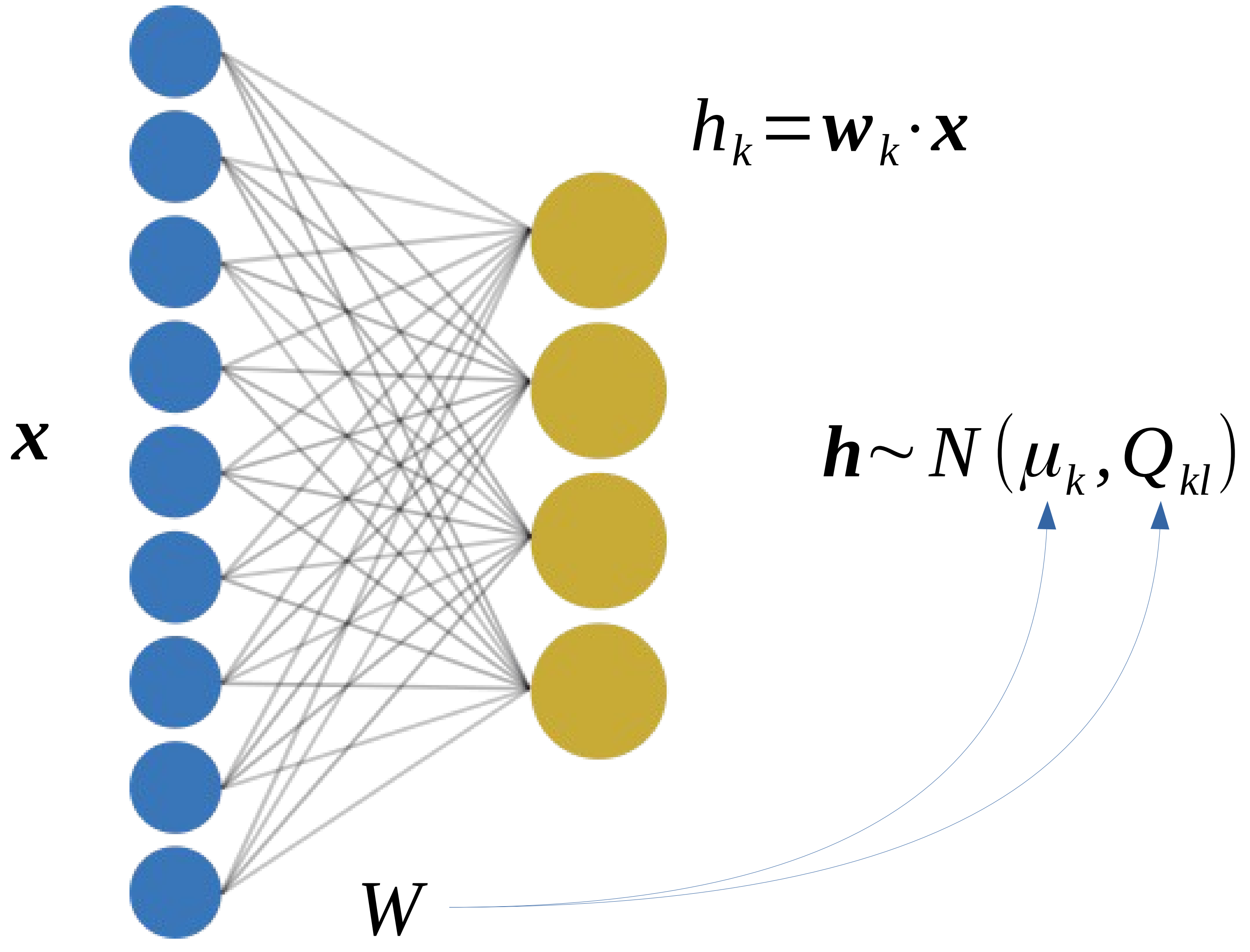
Alignment on "principal components" of **higher-order** image statistics → RF localization

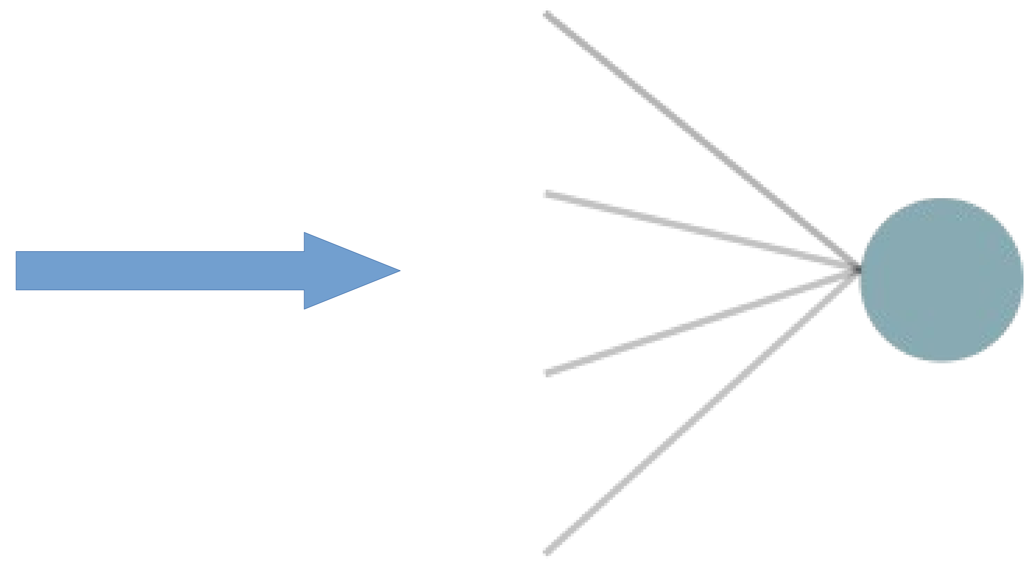


Gaussian equivalence



Gaussian equivalence





- Generalization error (prediction MSE, **pmse**)
- dynamics of gradient descent in terms of order parameters **Q**

 # parameters:

$DK \rightarrow K + K(K-1)/2$

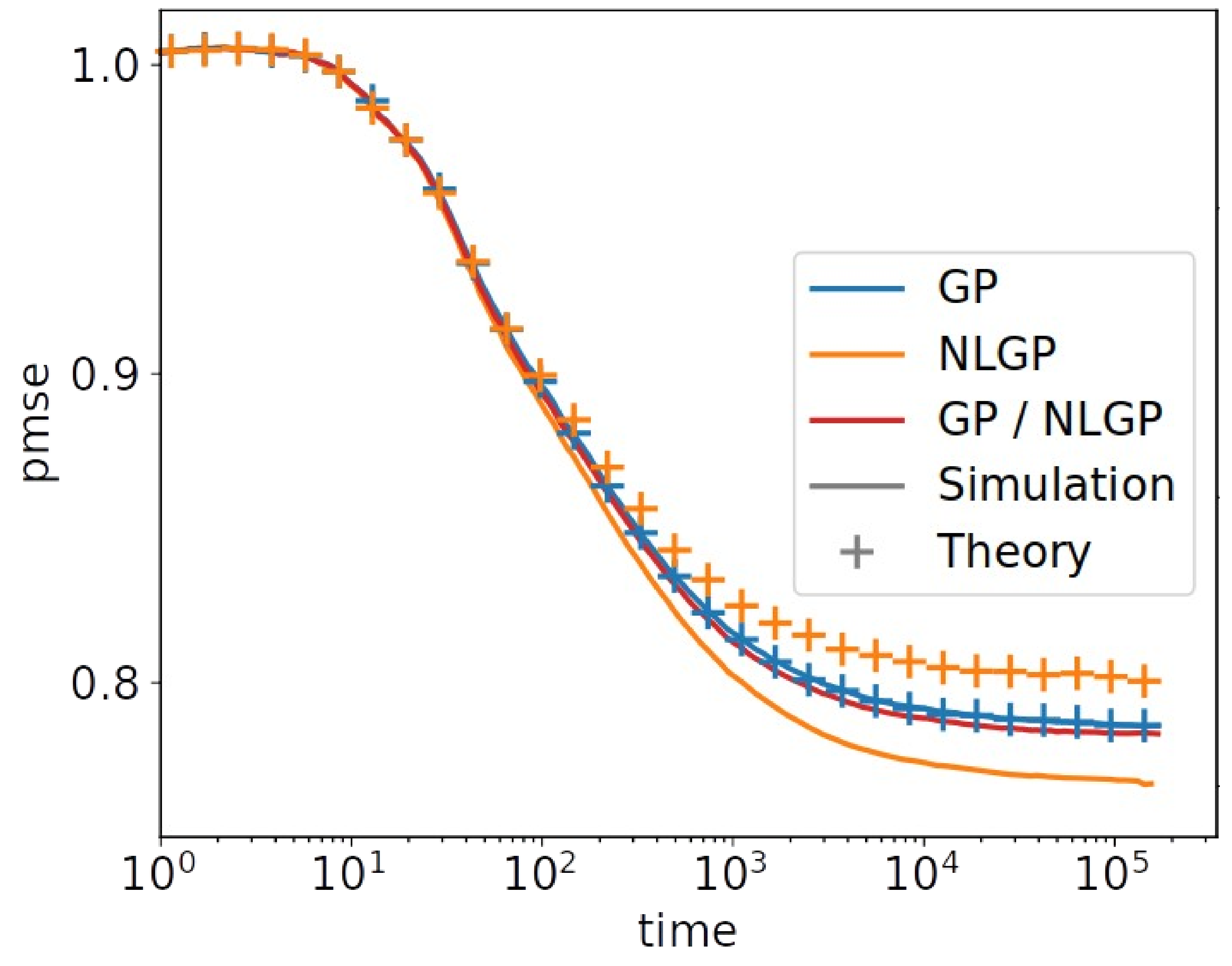
The limits of Gaussian equivalence (GET)

The formation of localised RF is not captured by an equivalent Gaussian model

Can we predict the loss?

- Can evaluate test error on Gaussian inputs analytically (Refinetti et al, ICML ’21)
- For NLGP inputs, need the GET (Goldt et al., MSML ’21) but **the GET breaks down**.

Legend:
trained on / tested on



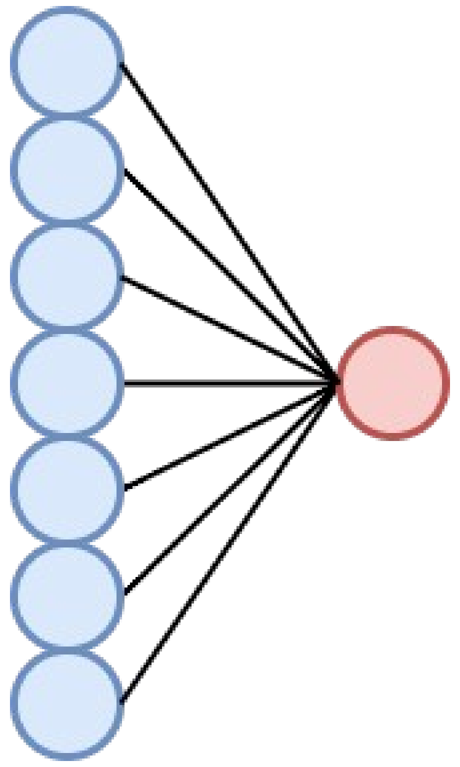
SO WHAT?

SINGLE NEURON,

OF COURSE

Connecting receptive fields to data geometry

A simplified model highlights the importance of non-Gaussian statistics



$$y = \sigma(w \cdot x)$$

$$\sigma(h) = \alpha h - \frac{\beta}{3} h^3$$

Gradient Flow (GF) dynamics:

$$\dot{\boldsymbol{w}} = \frac{1}{M} \sum_{\mu=1}^{M} \left(c_2^\mu C^\mu \boldsymbol{w} + c_4 T^\mu \boldsymbol{w}^{\otimes 3} \right)$$

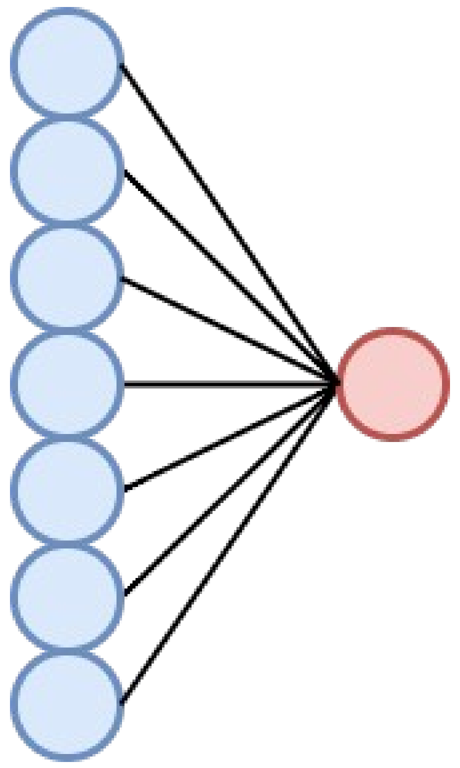
$$C_{ij}^\mu = \left\langle x_i^\mu x_j^\mu \right\rangle$$ same for **GP** and **NLGP** by construction

$$T_{ijk\ell}^\mu = \left\langle x_i^\mu x_j^\mu x_k^\mu x_\ell^\mu \right\rangle$$ split in Gaussian and non-Gaussian contribution

$$= C_{ij}^\mu C_{k\ell}^\mu + C_{ik}^\mu C_{j\ell}^\mu + C_{i\ell}^\mu C_{jk}^\mu + \Delta T_{ijk\ell}^\mu$$

Connecting receptive fields to data geometry

A simplified model highlights the importance of non-Gaussian statistics



$$y = \sigma(w \cdot x)$$

$$\sigma(h) = \alpha h - \frac{\beta}{3} h^3$$

Gradient Flow (GF) dynamics:

$$\dot{\boldsymbol{w}} = \frac{1}{M} \sum_{\mu=1}^{M} (c_2^\mu + c_4 q^\mu) C^\mu \boldsymbol{w} + \frac{c_4}{M} \sum_{\mu=1}^{M} \Delta T^\mu \boldsymbol{w}^{\otimes 3}$$

$$q^\mu = \boldsymbol{w}^T C^\mu \boldsymbol{w}$$

$$C_{ij}^\mu = \left\langle x_i^\mu x_j^\mu \right\rangle$$

same for **GP** and **NLGP** by construction

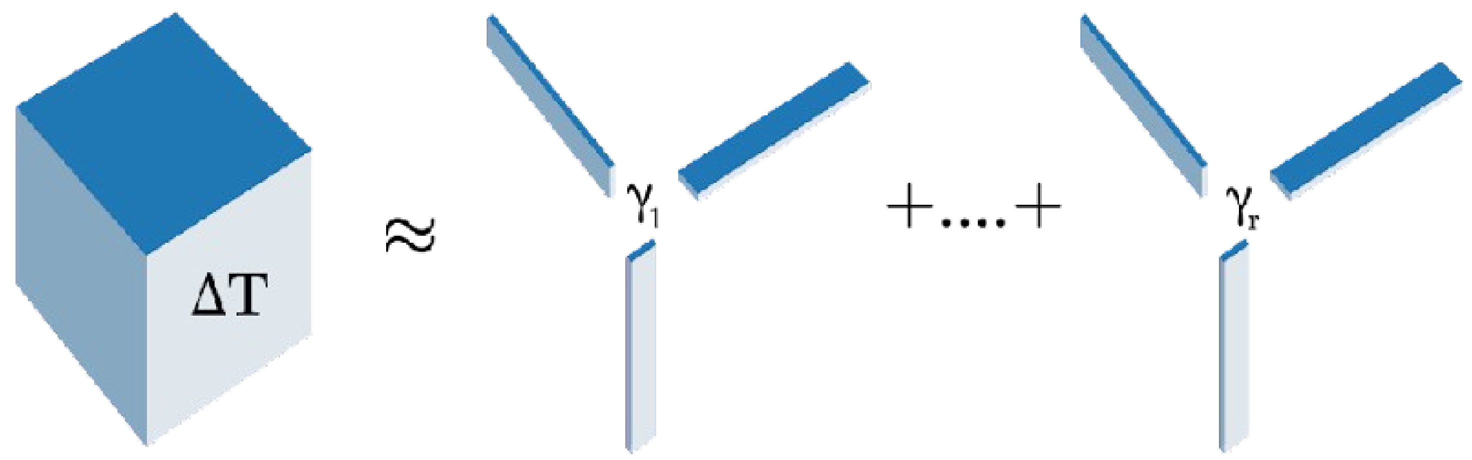
$$T_{ijk\ell}^\mu = \left\langle x_i^\mu x_j^\mu x_k^\mu x_\ell^\mu \right\rangle$$

split in Gaussian and non-Gaussian contribution

$$= C_{ij}^\mu C_{k\ell}^\mu + C_{ik}^\mu C_{j\ell}^\mu + C_{i\ell}^\mu C_{jk}^\mu + \Delta T_{ijk\ell}^\mu$$

Decomposing the 4th-order cumulant

$$\dot{\boldsymbol{w}} = \frac{1}{M}\sum_{\mu=1}^{M}\left(c_2^{\mu} + c_4 q^{\mu}\right)C^{\mu}\boldsymbol{w} + \frac{c_4}{M}\sum_{\mu=1}^{M}\Delta T^{\mu}\boldsymbol{w}^{\otimes 3}$$

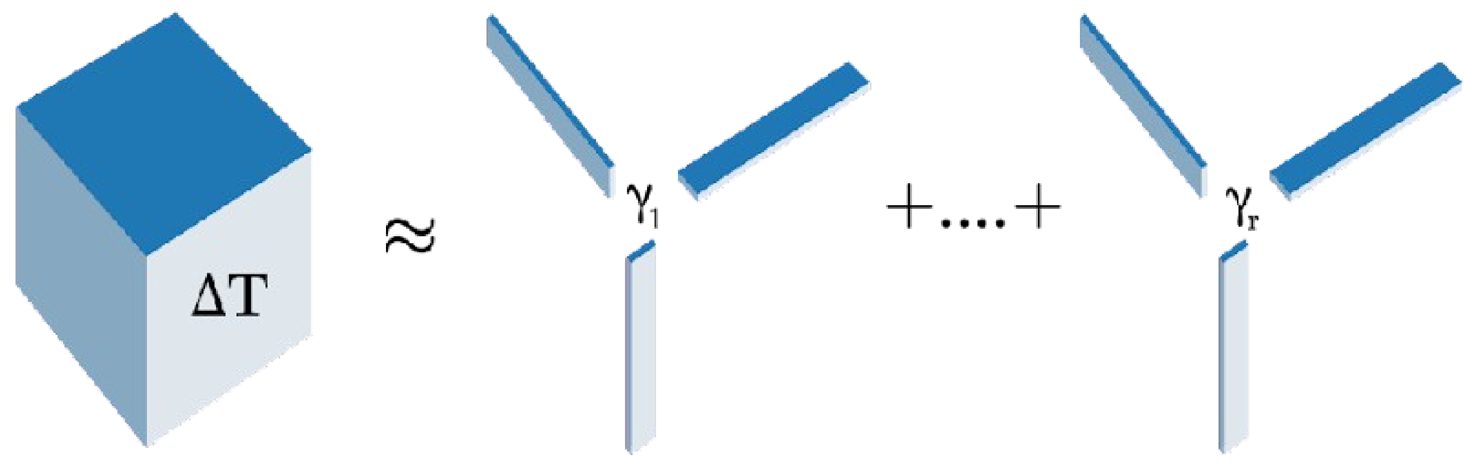


CP decomposition:

$$\Delta T = \sum_{k=1}^{r}\gamma_k \boldsymbol{u}_k \otimes \boldsymbol{u}_k \otimes \boldsymbol{u}_k \otimes \boldsymbol{u}_k$$

Decomposing the 4th-order cumulant

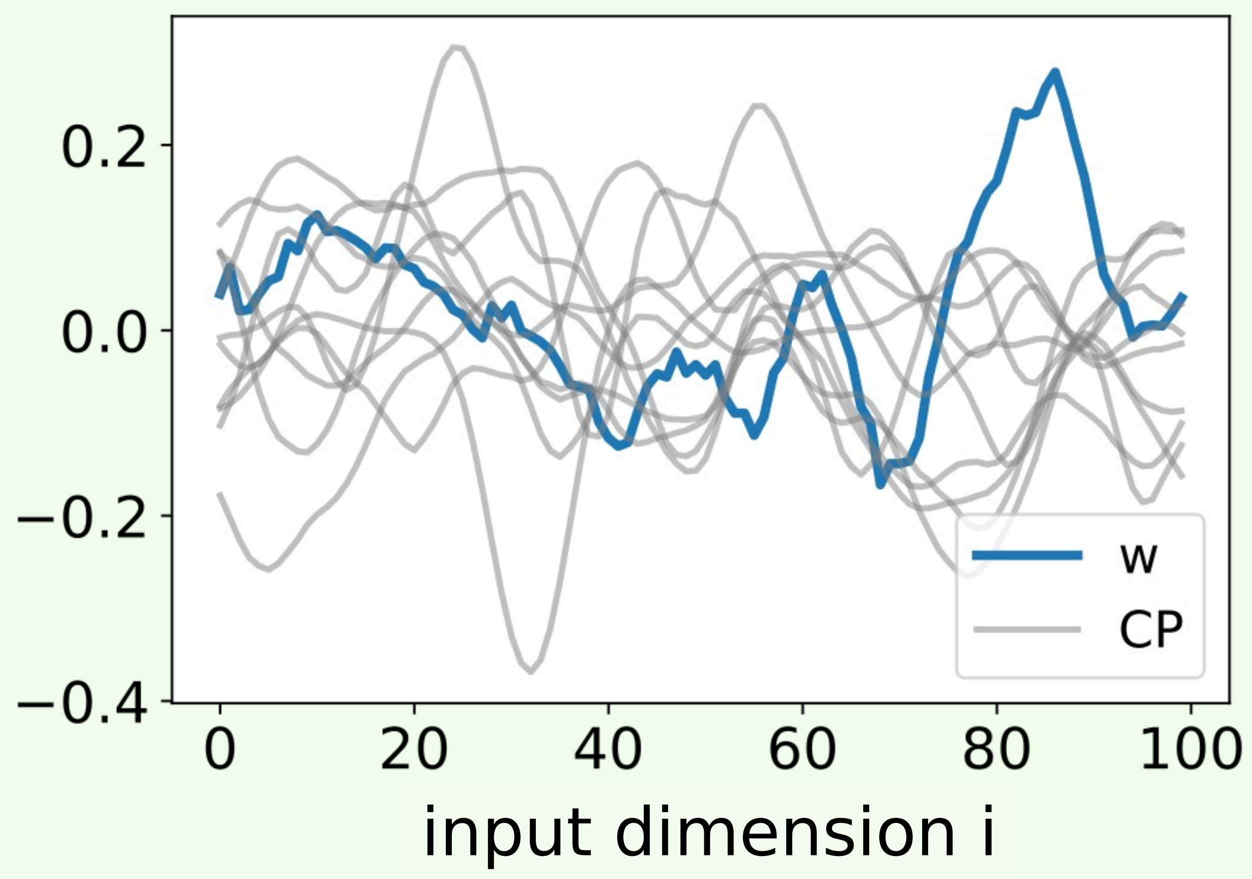
$$\dot{\boldsymbol{w}} = \frac{1}{M}\sum_{\mu=1}^{M}(c_2^\mu + c_4 q^\mu)C^\mu \boldsymbol{w} + \frac{c_4}{M}\sum_{\mu=1}^{M}\Delta T^\mu \boldsymbol{w}^{\otimes 3}$$



CP decomposition: $$\Delta T = \sum_{k=1}^{r}\gamma_k \boldsymbol{u}_k \otimes \boldsymbol{u}_k \otimes \boldsymbol{u}_k \otimes \boldsymbol{u}_k$$

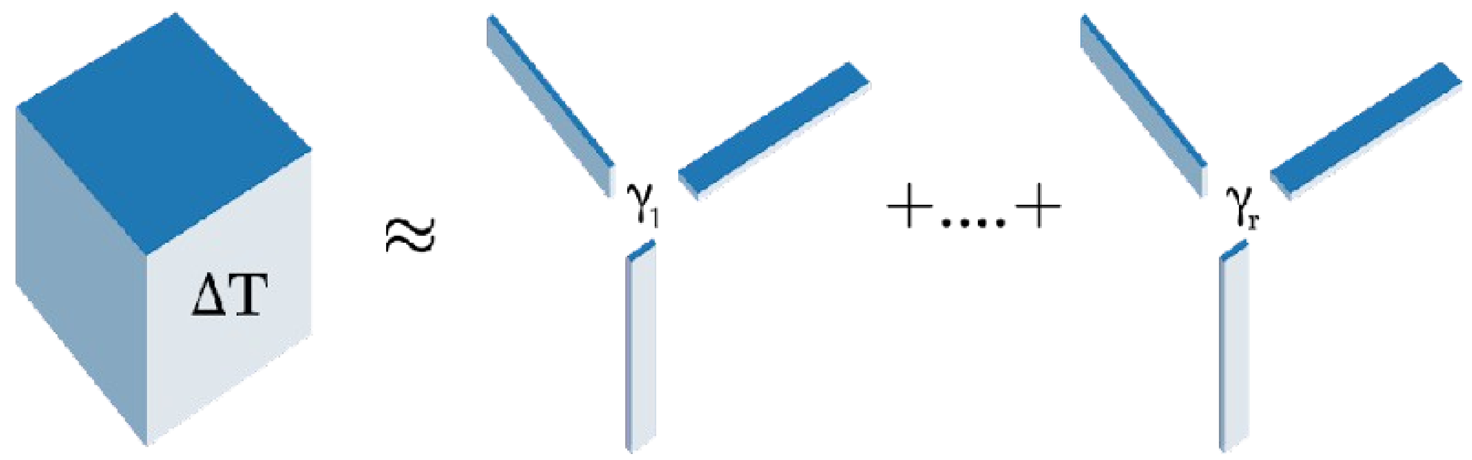
≈ Gaussian inputs

- CP factors oscillate
- weight vector is not localised



Decomposing the 4th-order cumulant

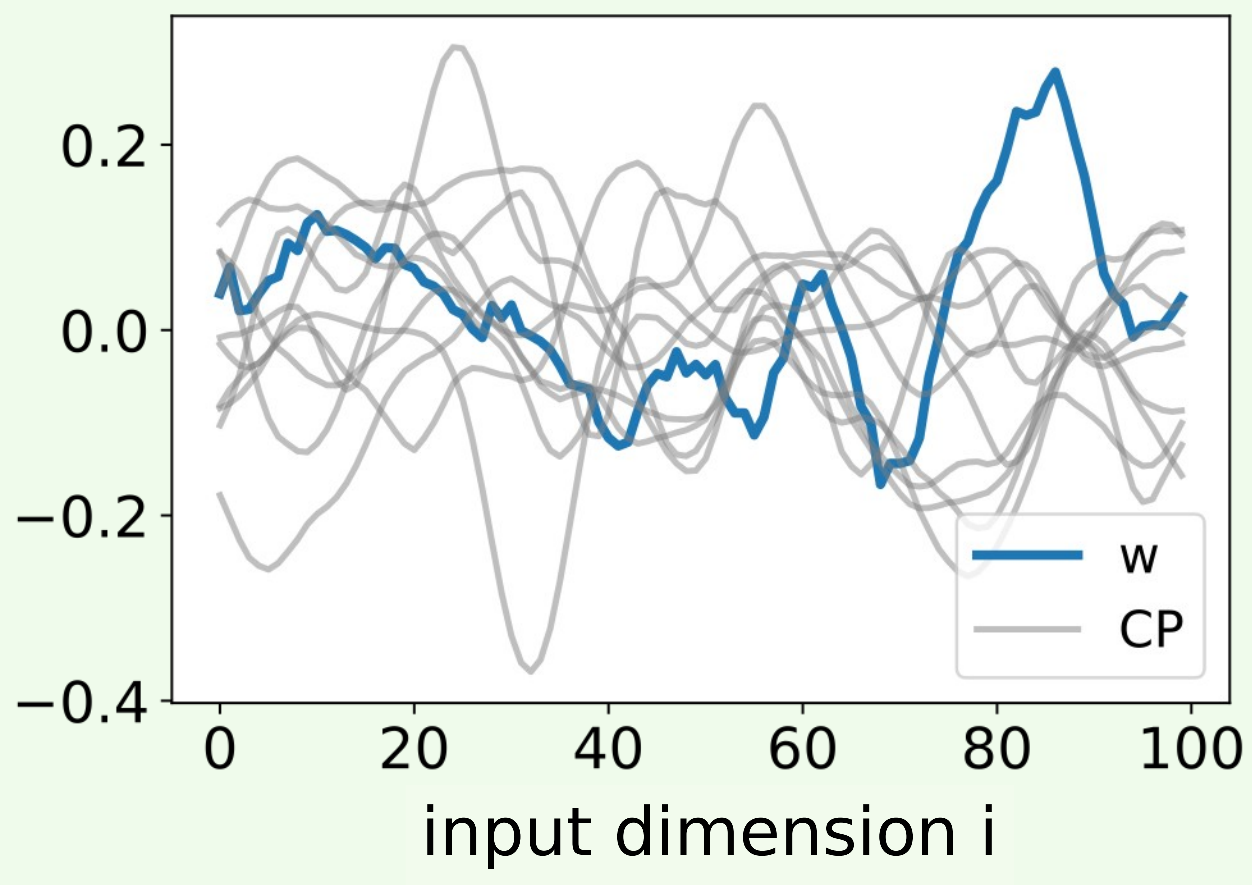
$$\dot{\boldsymbol{w}} = \frac{1}{M}\sum_{\mu=1}^{M}(c_2^\mu + c_4 q^\mu) C^\mu \boldsymbol{w} + \frac{c_4}{M}\sum_{\mu=1}^{M}\Delta T^\mu \boldsymbol{w}^{\otimes 3}$$



CP decomposition: $\Delta T = \sum_{k=1}^{r} \gamma_k \boldsymbol{u}_k \otimes \boldsymbol{u}_k \otimes \boldsymbol{u}_k \otimes \boldsymbol{u}_k$

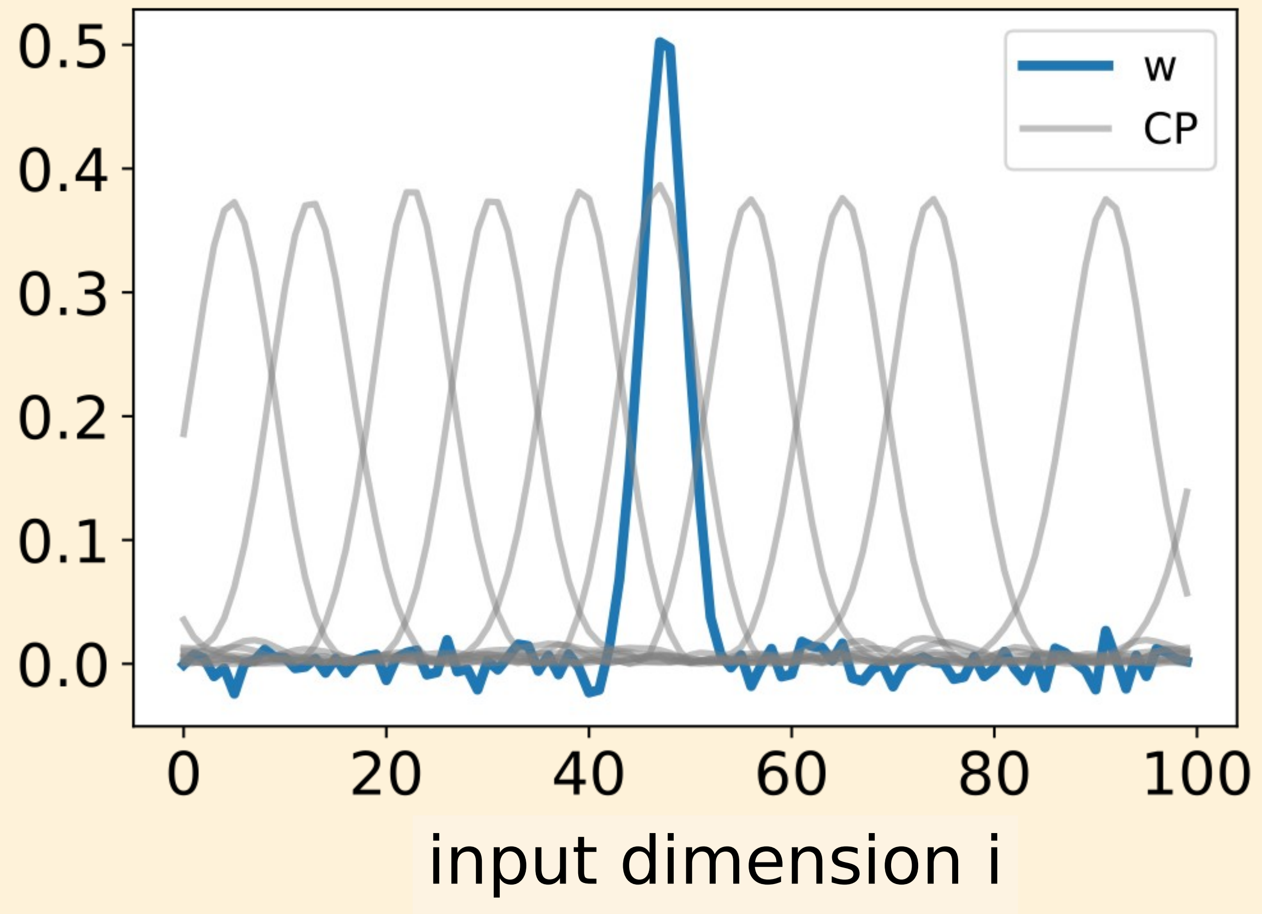
≈ Gaussian inputs

- CP factors oscillate
- weight vector is not localised



Non-gaussian inputs

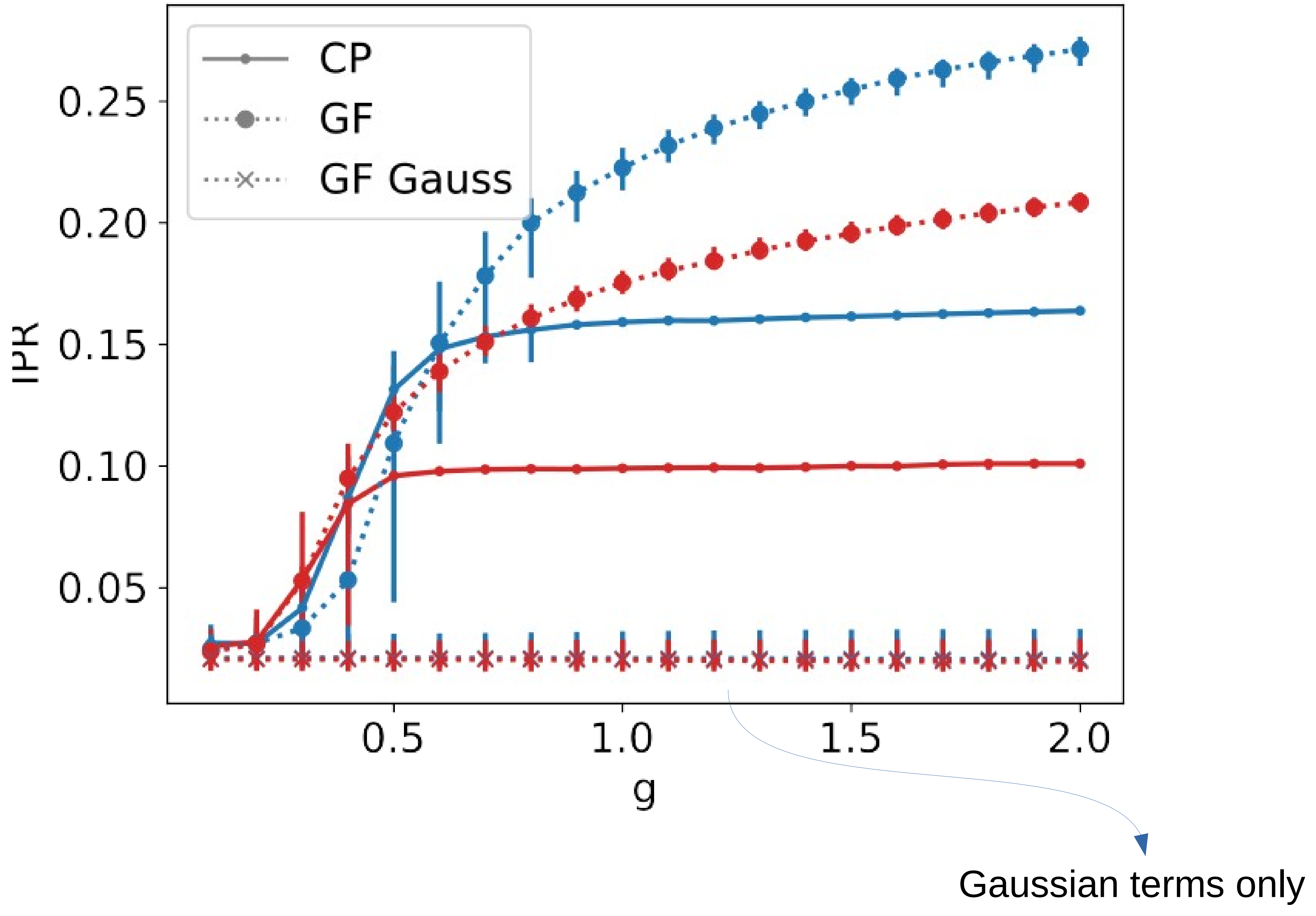
- CP factors localise
- and so does the weight vector



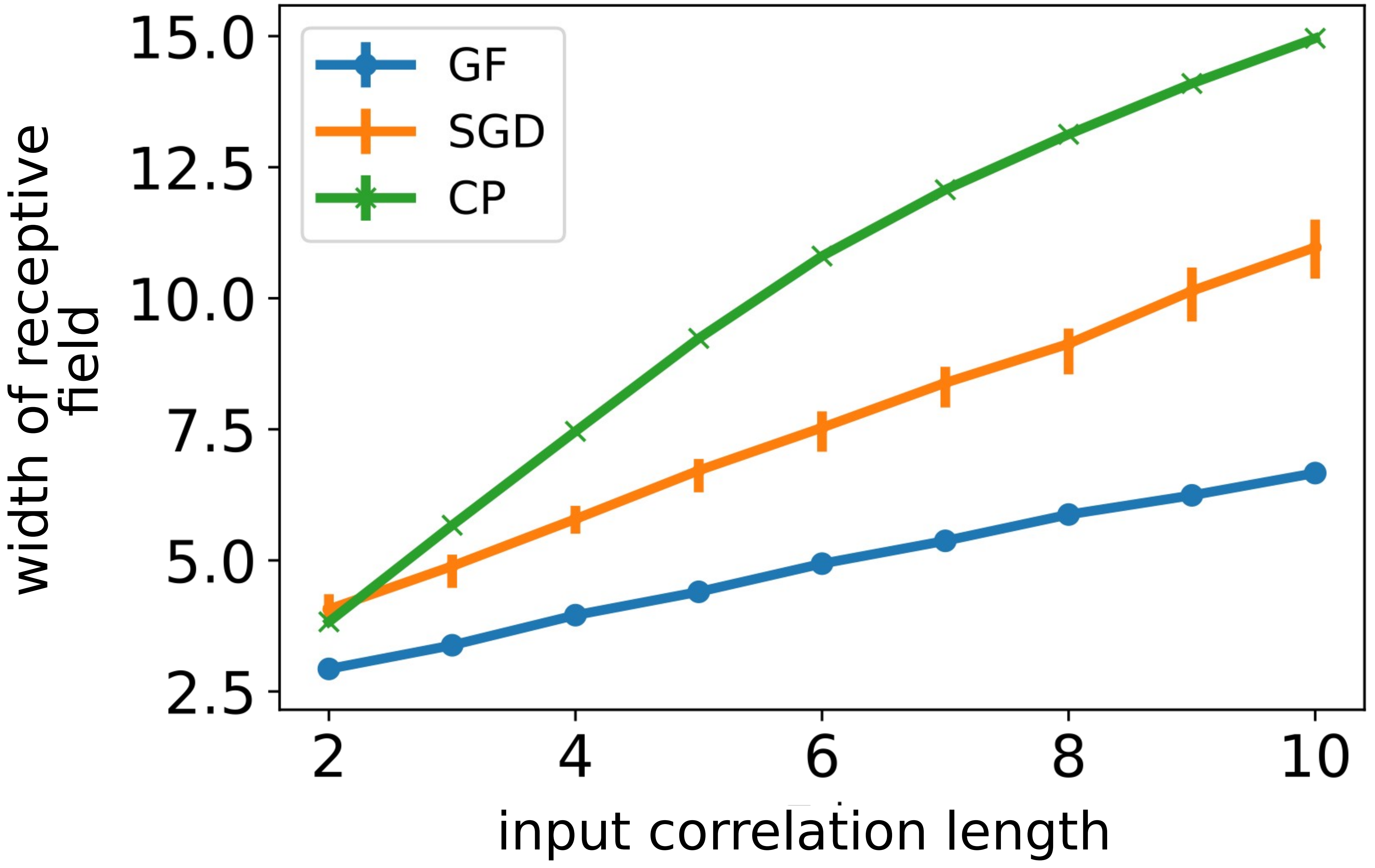
Relating cumulants and weight vectors

CP factors localisation → weight vector localisation

Localisation of receptive fields increases with gain:

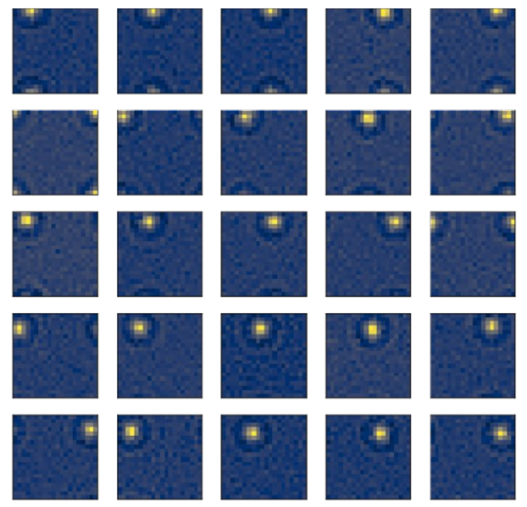


Width of CP factors ∼ width of the receptive field:

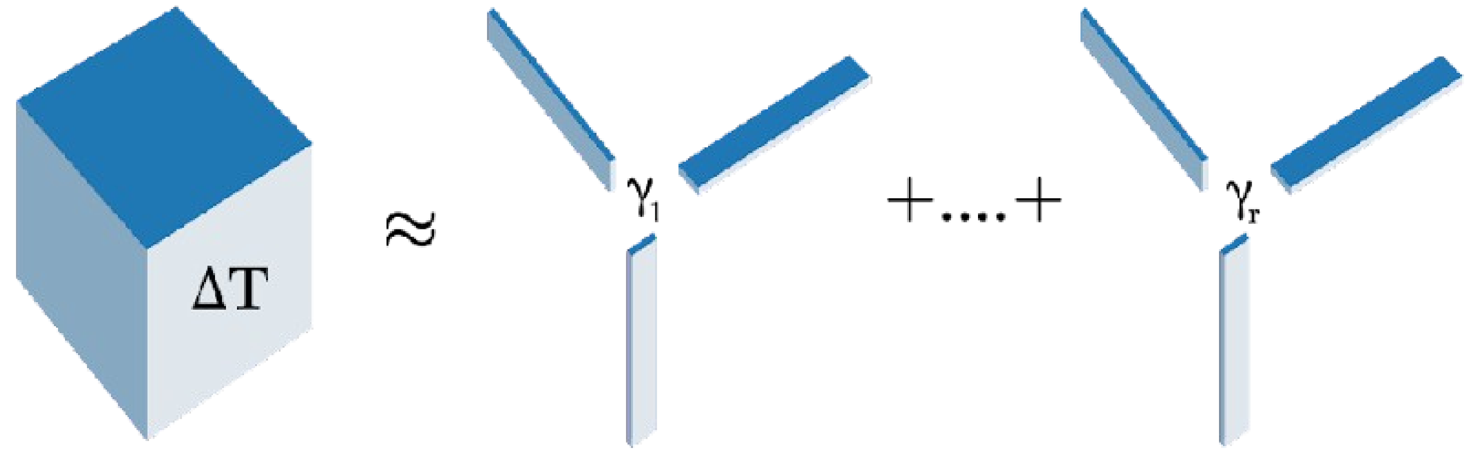


Concluding perspectives

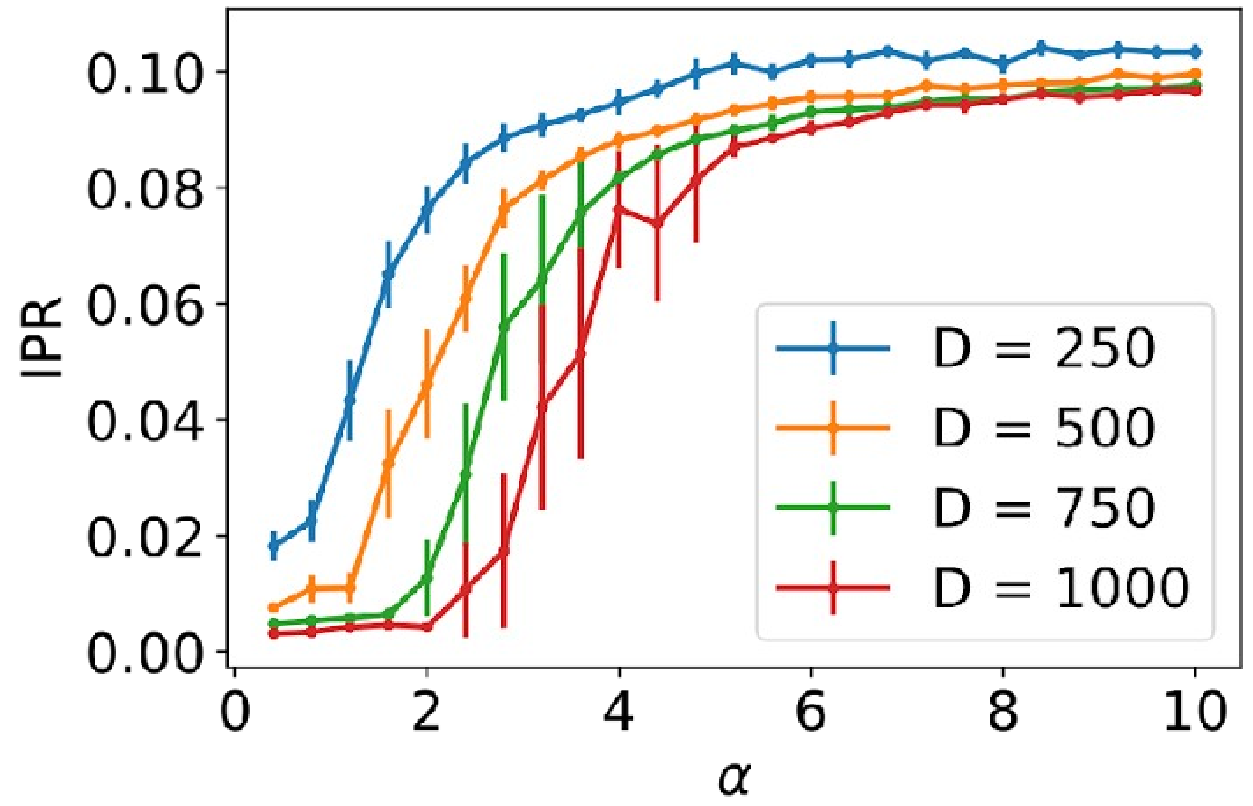
Going beyond Gaussian models for data



- Fully connected networks can learn a convolutional structure given the right statistical cues in their training data.



- Need better understanding of interaction btw higher-order tensors and learning dynamics.
  - Unsupervised learning: Harsh et al. ’20, Ocker & Buice ’21



- Transitions in higher-order random tensors.
- Impact of a general symmetry group on higher-order statistics → SGD learning dynamics.

THE END