PHYSICS OF AI
FROM ARTIFICIAL TO CORTICAL NETWORKS

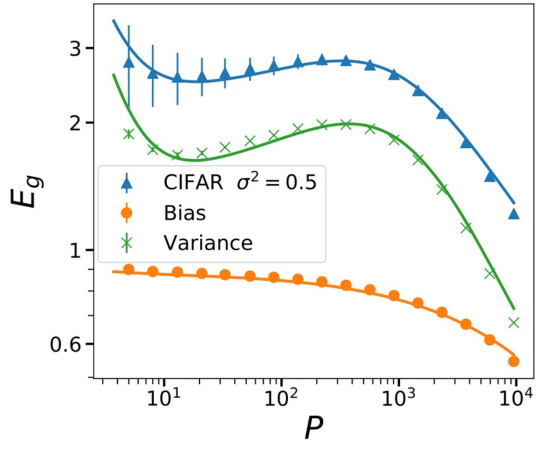
MORITZ HELIAS

THEORETICAL NEUROSCIENCE (IAS-6)

FACULTY OF PHYSICS, RWTH AACHEN UNIVERSITY

2024-10-02 CRC NUMERIQS BONN

PHYSICS OF AI



Canatar et al. (2021)

1. Learning and generalization

goals:
- optimal parameters / architectures?

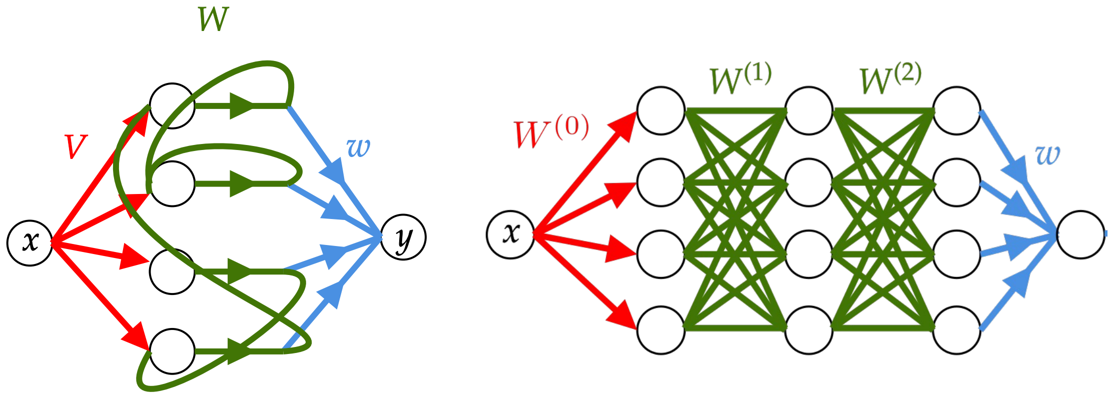
- implicit bias

  * why are deep networks (transformers) good architectures; are there better ones?

  * understand abilities and limitations of particular architectures

- sample efficiency:
  * how much data required to reach desired performance → energy demand

2. Link between biological and artificial neuronal networks



goals:
- understand qualitative similarities and differences

- identify features of biological networks that are essential for efficiency

- propose biologically-inspired paradigms of computation

PHYSICS APPROACH TO NEURONAL NETWORKS

Kuehn, Helias (2018)
J Phys A

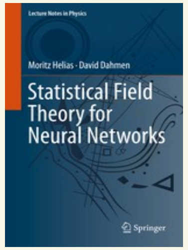
Helias, Dahmen (2020)
Springer Lecture Notes in Physics

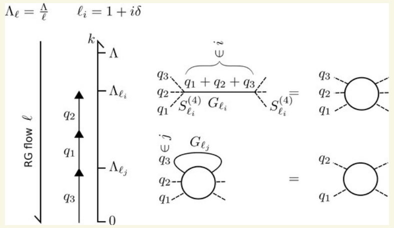
Stapmanns et. al. (2020)
Phys Rev E

Helias (2020)
J Phys A

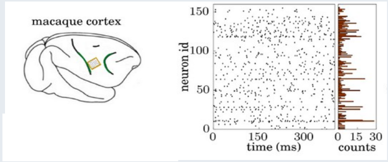
van Meegen, Kuehn, Helias (2021)
Phys Rev Lett

methods from physics

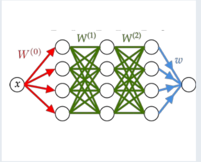


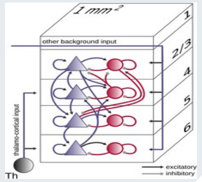


neuronal data



network models



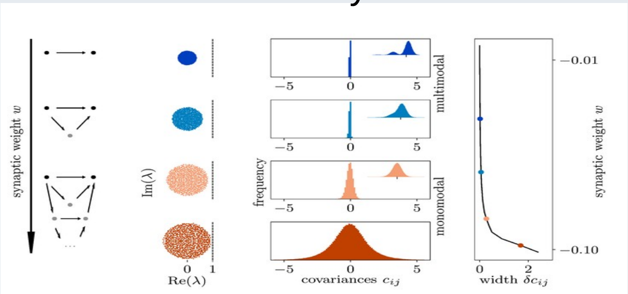


extract

analyze

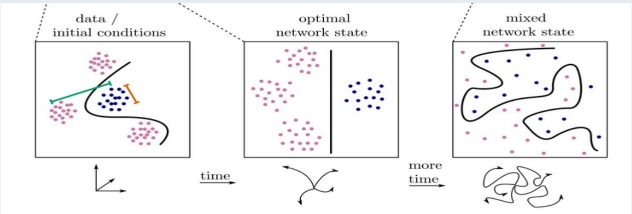
analyze

mechanisms of network dynamics



implement

neural information processing



input to

theory of AI, neuromorphic computing

Dahmen, Bos, Helias (2016) ***Phys Rev X***
Rostami, Porta-Mana, Gruen, Helias (2017) ***PLoS CB***
Dahmen, Gruen, Diesmann, Helias (2019) ***PNAS***
Senk et al. (2020) ***Phys Rev Res***
Layer et al. (2022) ***Elife***
Dahmen et al. (2024) ***PRX Life***

Schuecker, Goedeke, Helias (2018) ***Phys Rev X***
Dahmen, Gilson, Helias (2020) ***J Phys A***
Gilson, Moreno-Bote, Insabato, Dahmen, Helias (2020) ***PloS CB***
Nestler, Keup, Dahmen, Gilson, Rauhut, Helias (2020) ***NeurIPS***
Keup, Kuehn, Dahmen, Helias (2021) ***Phys Rev X***
Fischer, Keup, Rene, Layer, Dahmen, Helias (2022) ***Phys Rev Res***
Tiberi, Stapmann, Kuehn, Dahmen, Luu, Helias *(2022)* ***Phys Rev Lett***
Merger, Rene, Fischer, Dahmen, Helias (2023) ***Phys Rev X***
Fischer, Lindner, Dahmen, Ringel, Kraemer, Helias (2024) ***ICML***
Epping, Rene, Helias, Schaub (2024) ***NeurIPS***

1. FROM BAYESIAN INFERENCE TO PHYSICS OF AI

2. FROM DEEP TO RECURRENT BIOLOGICAL NETWORKS

1. FROM BAYESIAN INFERENCE TO PHYSICS OF AI

SUPERVISED LEARNING AS BAYESIAN INFERENCE

Training data

$$\mathcal{D} = \{(x_\alpha, y_\alpha)\}_{1 \le \alpha \le P}$$

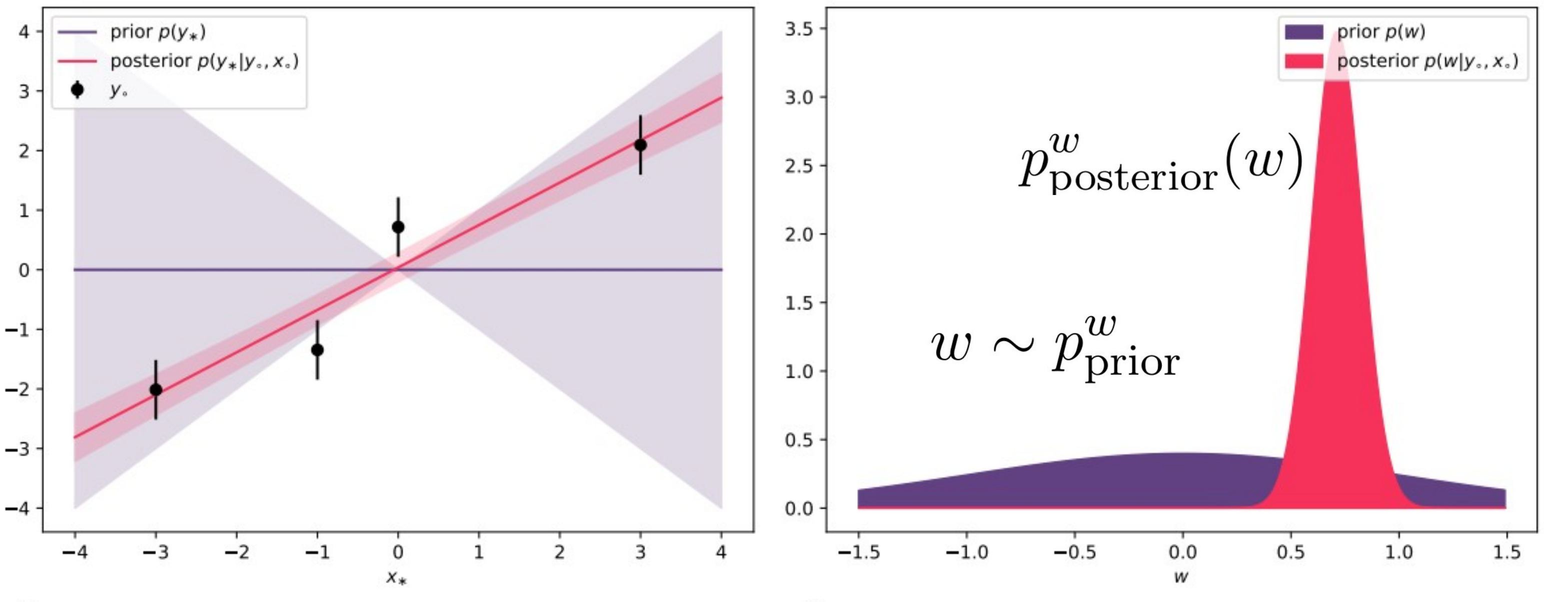
Model

$$y_\alpha = w \cdot x_\alpha$$

$$\rightarrow p(y|x, w)$$

Prior

$$w \sim p^w_{\mathrm{prior}}$$

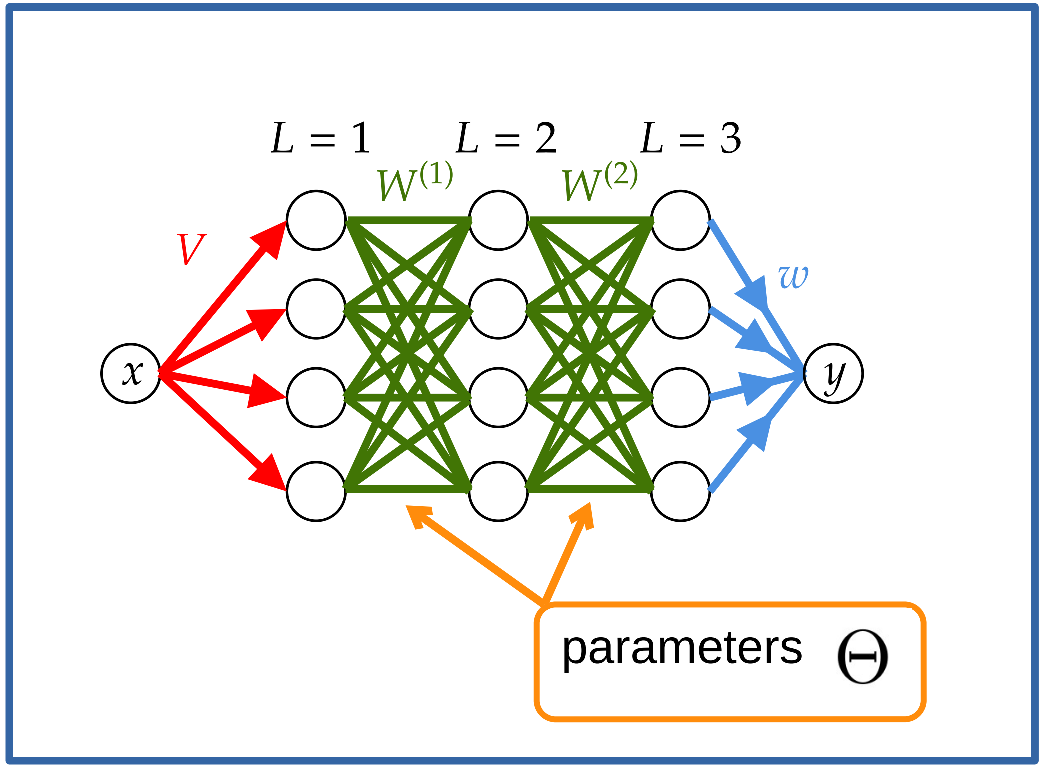


Posterior (Bayes' law)

$$p^w_{\mathrm{posterior}}(w) = \frac{p(y|X, w)\, p_{\mathrm{prior}}(w)}{p(y)}$$

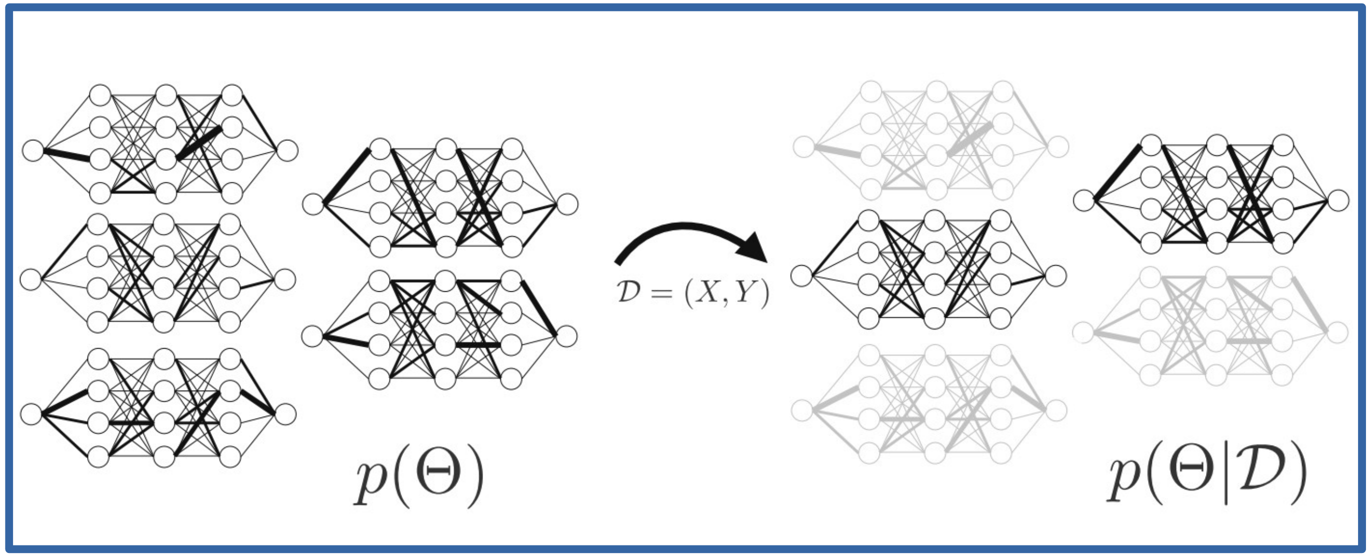
APPLICATION TO DEEP NETWORK

single network



prior ensemble of networks

posterior ensemble



only requires mapping by network

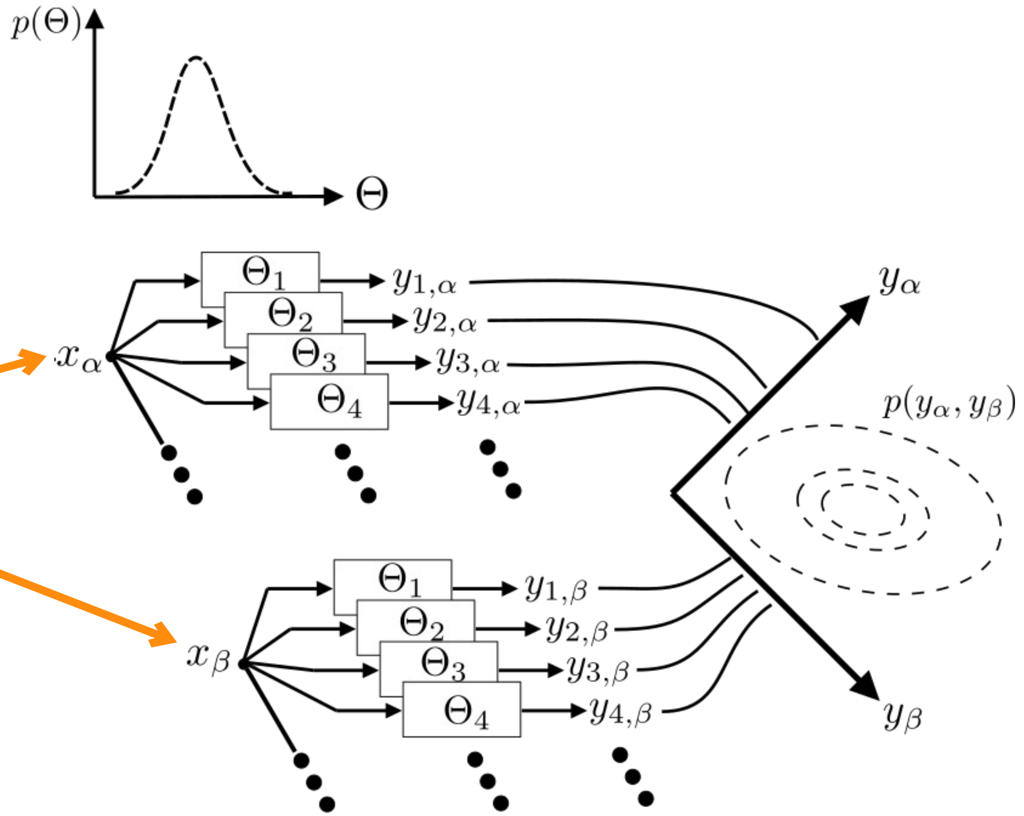
$$\Theta, X \mapsto y$$

weight posterior = equilibrium distribution
of **noisy gradient descent**
with weight regularization

$$\frac{\partial}{\partial t}\Theta = -\nabla_\Theta \left[\mathcal{L} + \|\Theta\|^2\right] + \text{noise}$$

APPLICATION TO DEEP NETWORK: PRIOR OF OUTPUTS

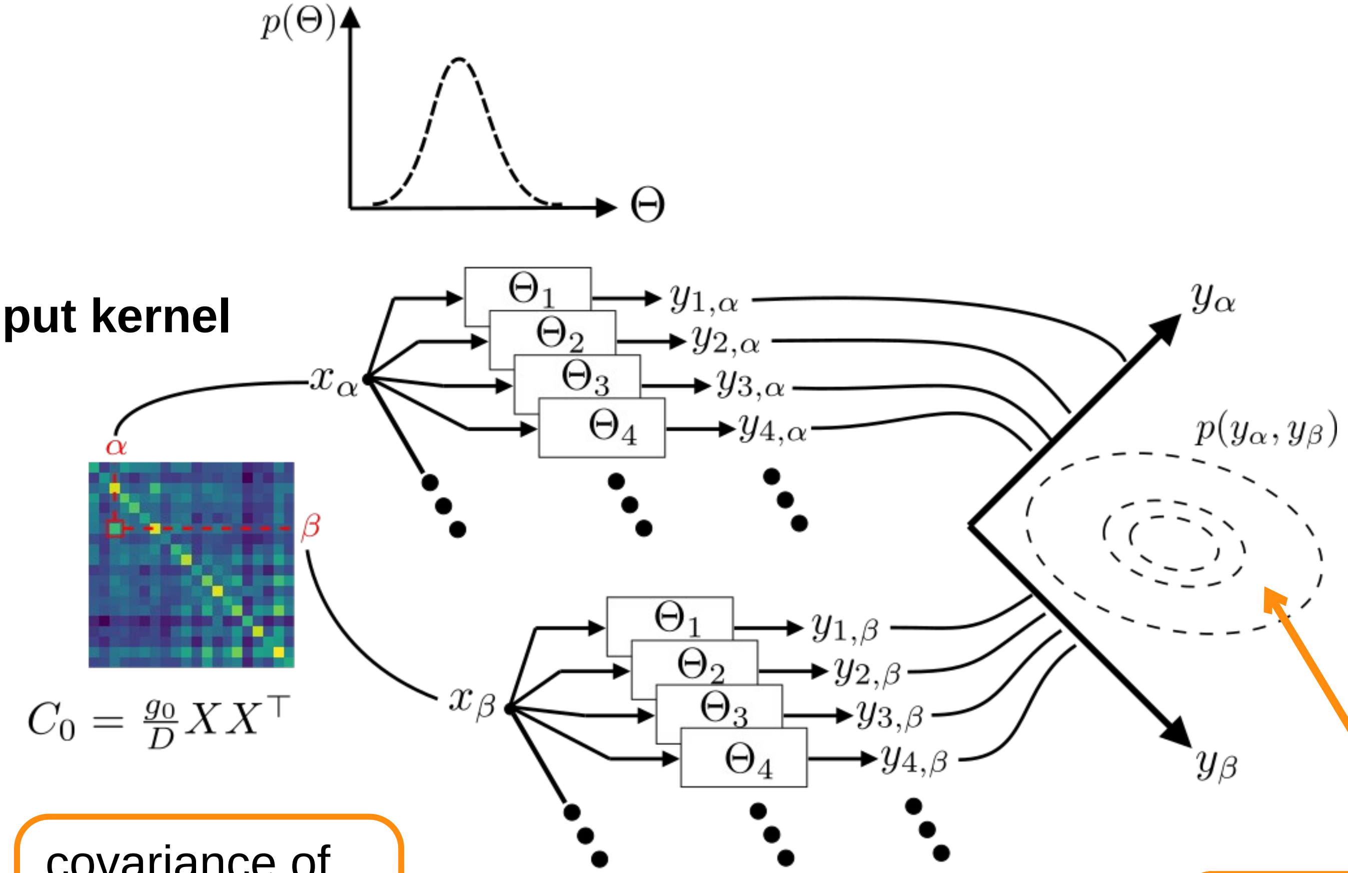
different input data points



joint prior distribution of network outputs

GAUSSIAN PROCESS THEORY OF LEARNING

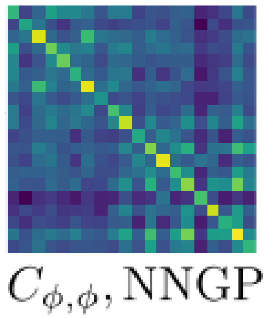
input kernel



covariance of inputs (similarity)

approximate as Gaussian

output kernel



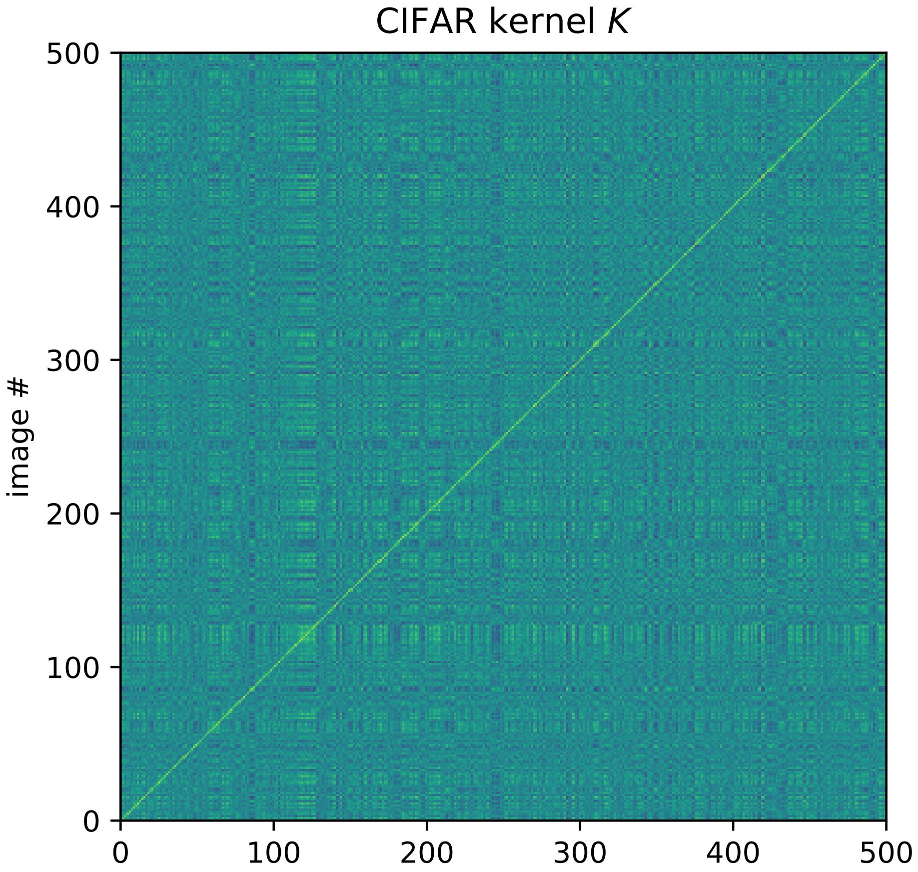
covariance of outputs (similarity)

Williams et al. (1998)

MEANING OF THE GAUSSIAN KERNEL: SIMILARITY

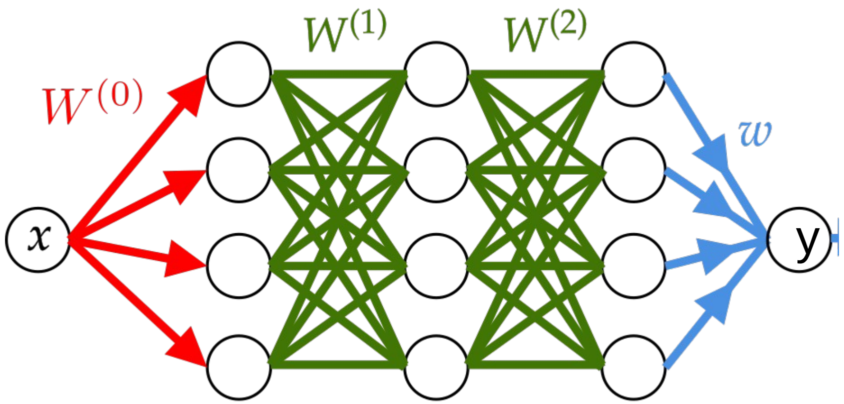
CIFAR-10 dataset

input kernel



$$C^{(XX)} = \frac{g_0}{D} X X^{\mathsf{T}}$$

SETTING: DEEP NETWORK



network mapping

$$h_\alpha^{(0)} = W^{(0)} x_\alpha$$

$$h_\alpha^{(l)} = W^{(l)} \phi(h_\alpha^{(l-1)})$$

$$h \in \mathbb{R}^N$$

width = N

output

$$y_\alpha = h_\alpha^{(L)}$$

training data

$$\mathcal{D} = \{(x_\alpha, y_\alpha)\}_{1 \le \alpha \le P}$$

inputs $X \in \mathbb{R}^{P \times D}$

outputs / labels $Y \in \mathbb{R}^P$

\# training samples = P

BAYESIAN INFERENCE: PRIOR OF OUTPUT

$$W_{ij}^{(0)} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, g_0/D\right)$$
$$W_{ij}^{(l)} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, g_l/N\right)$$

enforce network equations for all L layers and all samples α

expectation over prior of weights

network prior

$$p(h|X) = \left\langle \prod_{\alpha=1}^{P} \prod_{l=1}^{L} \delta\left[-h_\alpha^{(l)} + W^{(l)}\phi(h_\alpha^{(l-1)})\right] \right\rangle_{W^{(1,\ldots,L)}, h^{(0)} \sim \mathcal{N}(0, C^{(XX)})}$$

$$h_\alpha^{(0)} = W^{(0)} x_\alpha \quad \rightarrow \quad C^{(XX)} = \frac{g_0}{D} X X^\mathsf{T}$$

PRIOR: CONTINUOUS SUPERPOSITION OF GAUSSIANS

field theory

$$p(Y|X) = \int \mathcal{D}C \int \mathcal{D}\tilde{C}\, \exp\left(S(C,\tilde{C}|Y)\right)$$

action

$$S(C,\tilde{C}|Y) = \ln \mathcal{N}\left(Y|0, C^{(L)}\right) - \boxed{N}\left[\mathrm{tr}\, \tilde{C}^{\mathsf{T}} C + \mathcal{W}(\tilde{C}|C)\right]$$

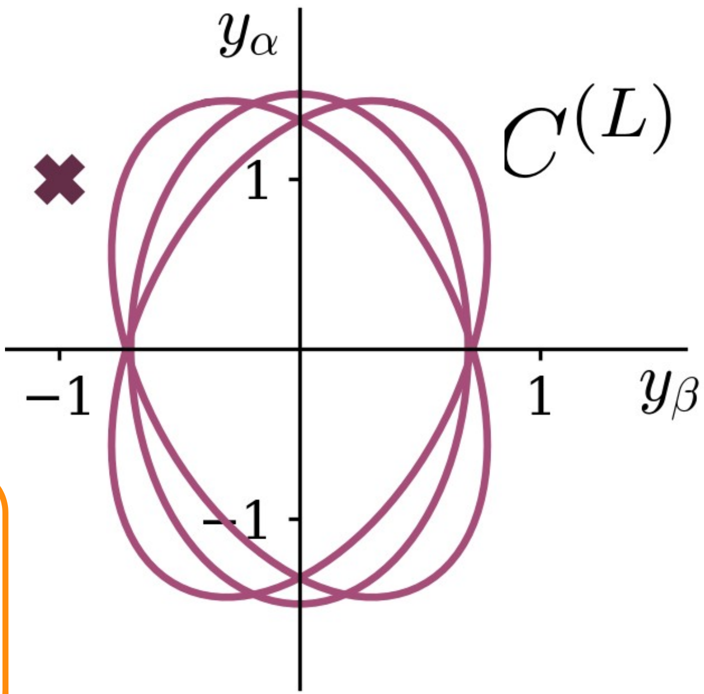
data term
$\propto P$

“fluctuating” kernel

$$C^{(l)}_{\alpha\beta} = \frac{g_l}{N} \phi^{(l-1)}_{\alpha} \cdot \phi^{(l-1)}_{\beta}$$

dominates integral in the limit of **large width N**

dot product over neurons
neurons within layer identical



Fischer et al. 2024 (ICML)

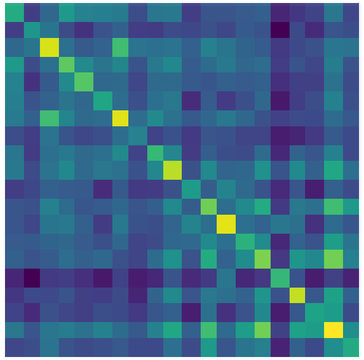
NEURAL NETWORK GAUSSIAN PROCESS (NNGP)

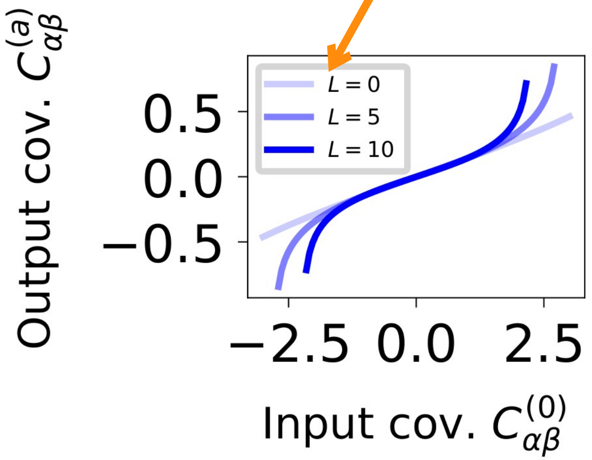
width $N \longrightarrow \infty$

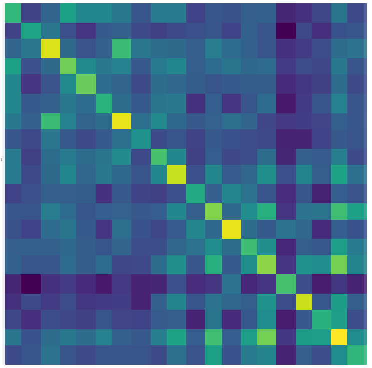
P finite

depth L

learning reduced to Gaussian process regression







input kernel

$$C^{(0)}_{\alpha\beta} = \frac{g_0}{D}\, x_\alpha \cdot x_\beta$$

output kernel

$$\{y_\alpha\} \sim \mathcal{N}(0, C^{(L)})$$

forward mapping of intermediate kernels

$$C^{(l+1)}_{\alpha\beta} = g_{l+1} \left\langle \phi(h_\alpha)\, \phi(h_\beta) \right\rangle_{\{h_\alpha\}\sim\mathcal{N}(0,C^{(l)})}$$

independent of targets

Williams (1998), Lee et al. (2018)

WHY GO FURTHER?

Lazy learning versus feature learning

There are two regimes in the theory of neural networks:

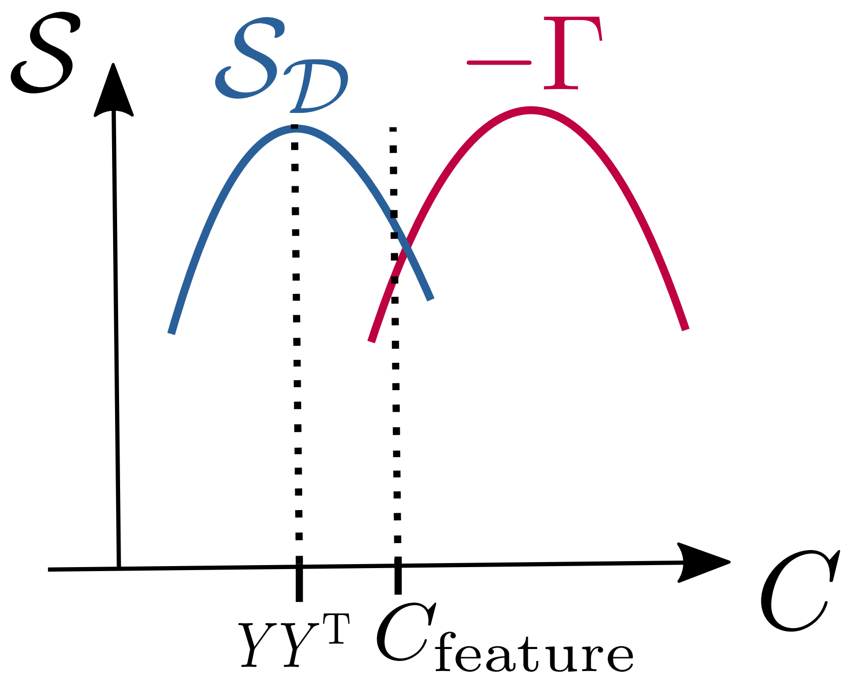
- lazy learning (Chizat et al., 2019)
  - → neural network Gaussian process (NNGP) (Neal 1994; Williams 1998; Lee et al., 2018)
    - equivalent to random feature regression (Mei et al., 2022)
  - → neural tangent kernel (NTK) (Jacot et al., 2018)
    - equivalent to linearization in weights

- **feature learning**
  - → network parameters adapt to task and network learns features of task
  - → networks typically show better performance (Geiger et al., 2020)
  - related works:
    Naveh & Ringel 2021; Zavatone-Veth, …, Pehlevan (2021);
    Li & Sompolinsky, 2021; Hanin & Zlokapa, 2023; Seroussi et al., 2023;
    Pacelli et al., 2023; Cui et al., 2023

FEATURE LEARNING

Proportional limit -> feature learning

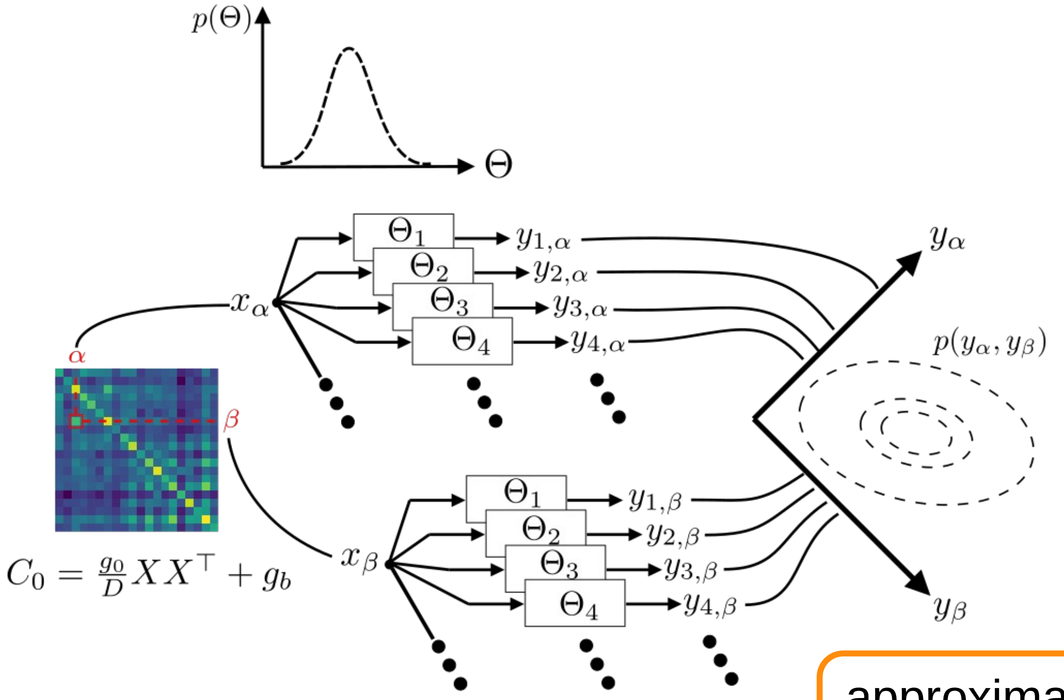
width	N	$\longrightarrow \infty$
# data points	$P = \alpha N$	$\longrightarrow \infty$



data term $\propto P$ net prior $\propto N$

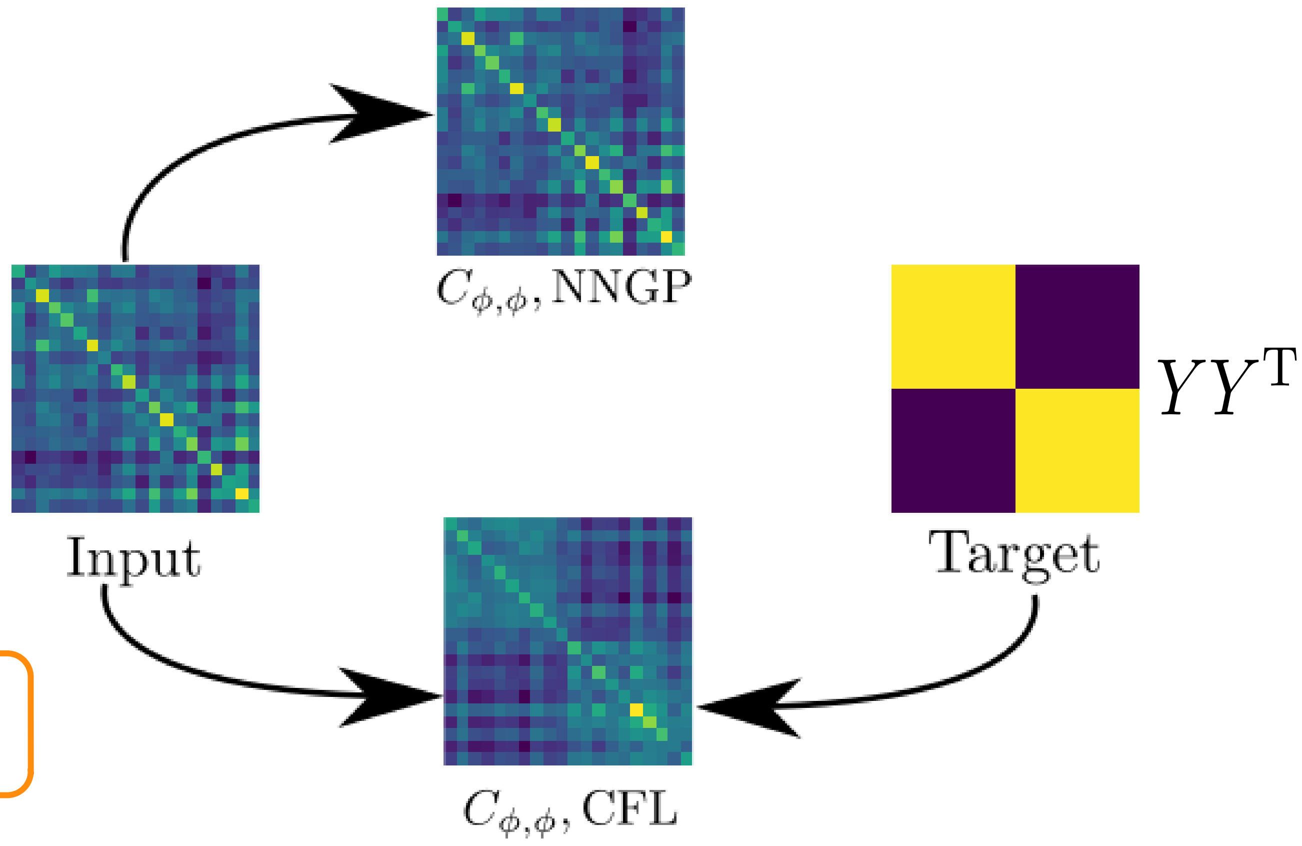
$$S(C,\tilde{C}|Y) = \underbrace{\ln \mathcal{N}\left(Y|0, C^{(L)}\right)}_{\mathcal{S}_{\mathrm{D}}(C^{(L)})} - \underbrace{N\left[\mathrm{tr}\,\tilde{C}^{\mathsf{T}}C + \mathcal{W}(\tilde{C}|C)\right]}_{\Gamma(C,\tilde{C})}$$

GAUSSIAN PROCESS THEORY OF LEARNING



approximate as Gaussian

Williams et al. 1998



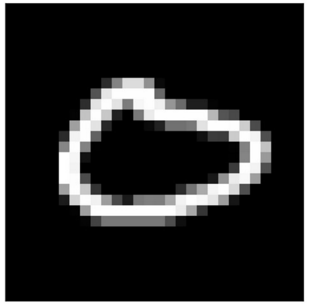
Fischer et al. 2024 (ICML)

RWTH AACHEN UNIVERSITY

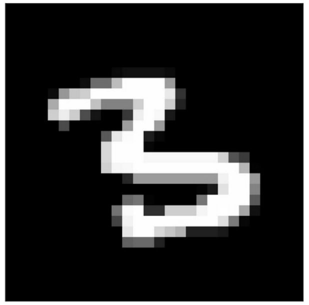
JÜLICH Forschungszentrum

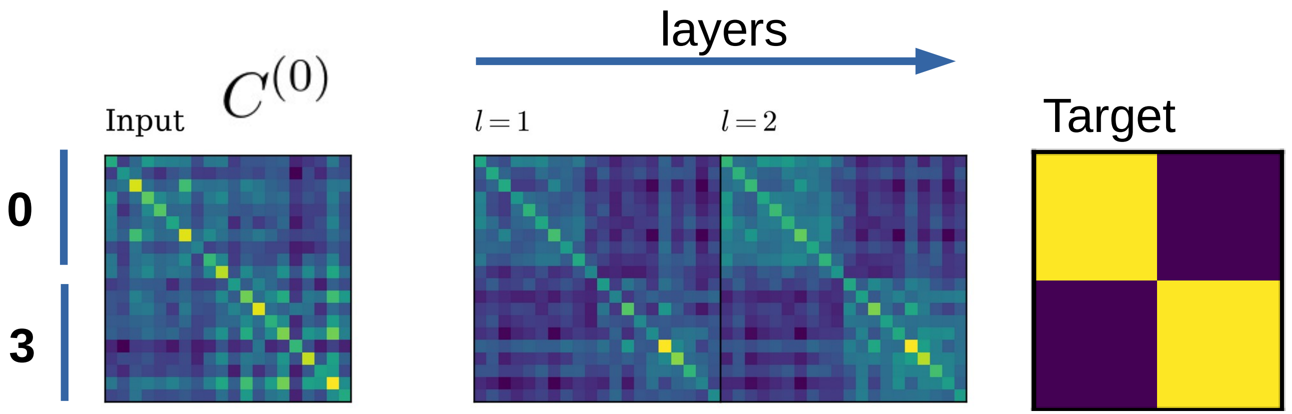
MNIST – CLASSIFICATION BETWEEN 0'S AND 3'S

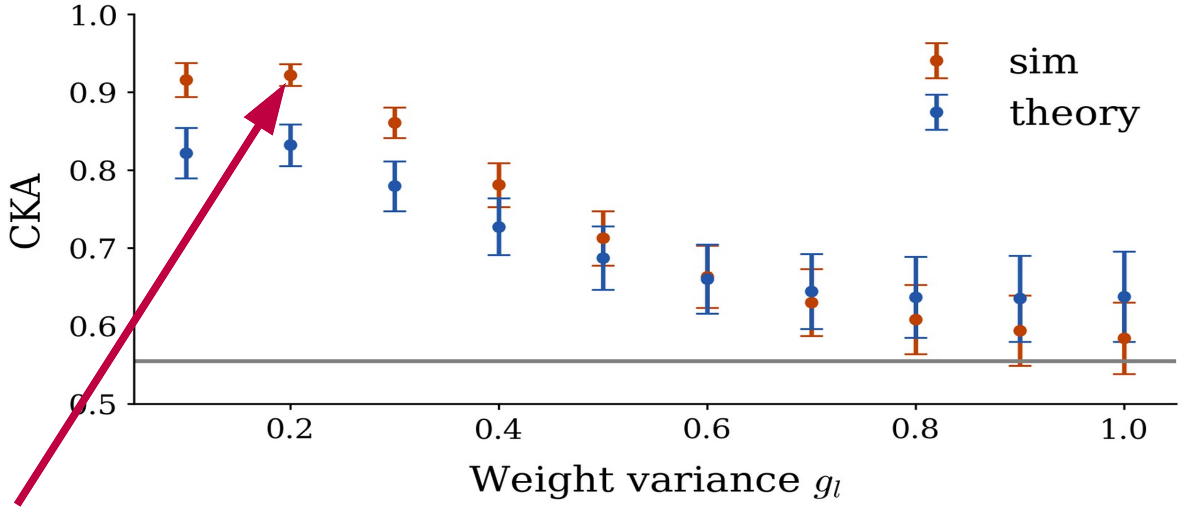
Numerical evaluation of theory



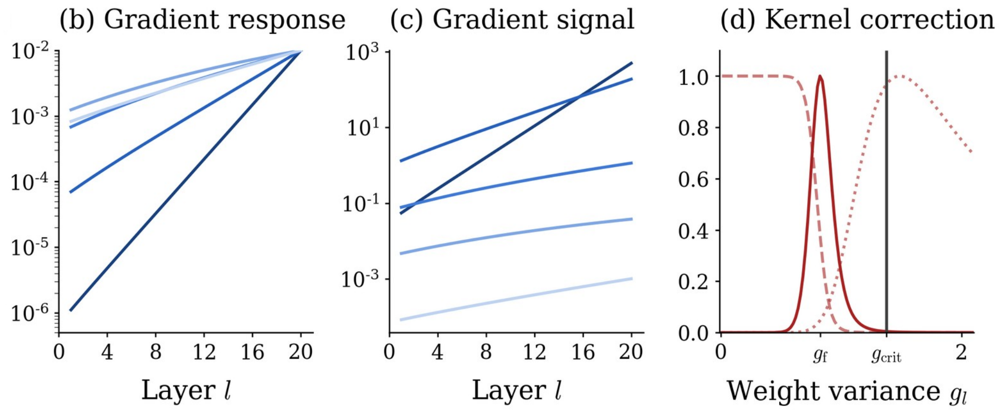
vs







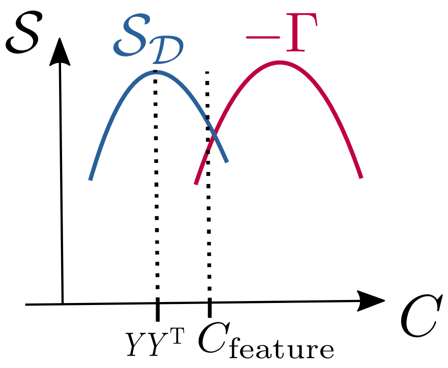
optimal alignment



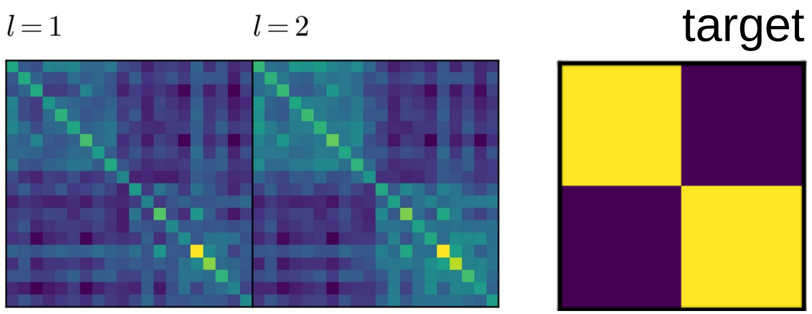
Fischer et al. 2024 (ICML)

INTERIM SUMMARY

- **Bayesian networks: prior is superposition of Gaussians**
  intermediate layers' kernels appear as order parameters

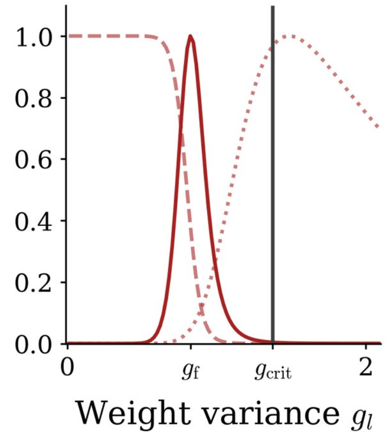


- **exact expressions for the Bayesian MAP kernels** in the proportional limit N, P → ∞ from saddle point of action



- **kernel adaptation in non-linear networks**
  discrepancy signal aligns kernel with target

- **tradeoff between critical fluctuations and output scale**
  shifts optimal adaptation towards smaller variance in weights

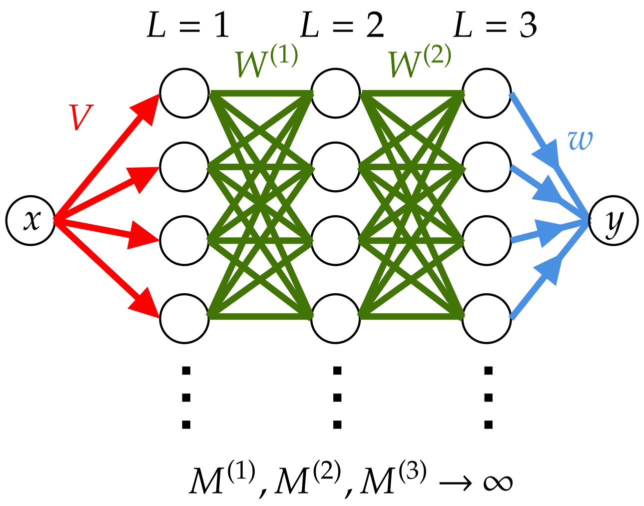


Fischer, Lindner et al. ICML 2024
arxiv 2405.10761

2. FROM DEEP ARTIFICIAL TO BIOLOGICAL NETWORKS

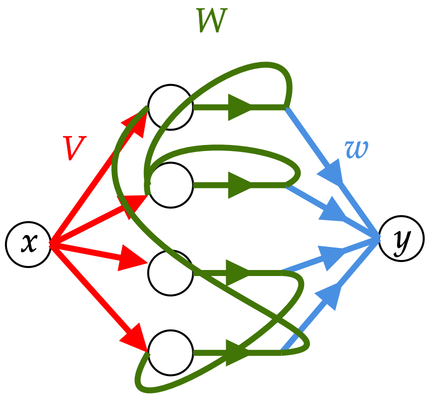
DEEP AND RECURRENT NETWORKS

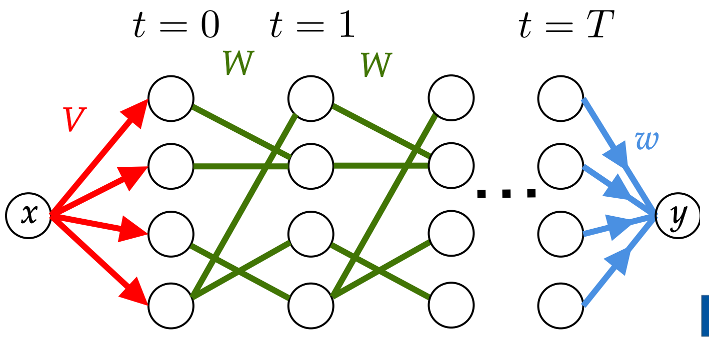
deep network



different $W^{(l)}$
between each pair of layers

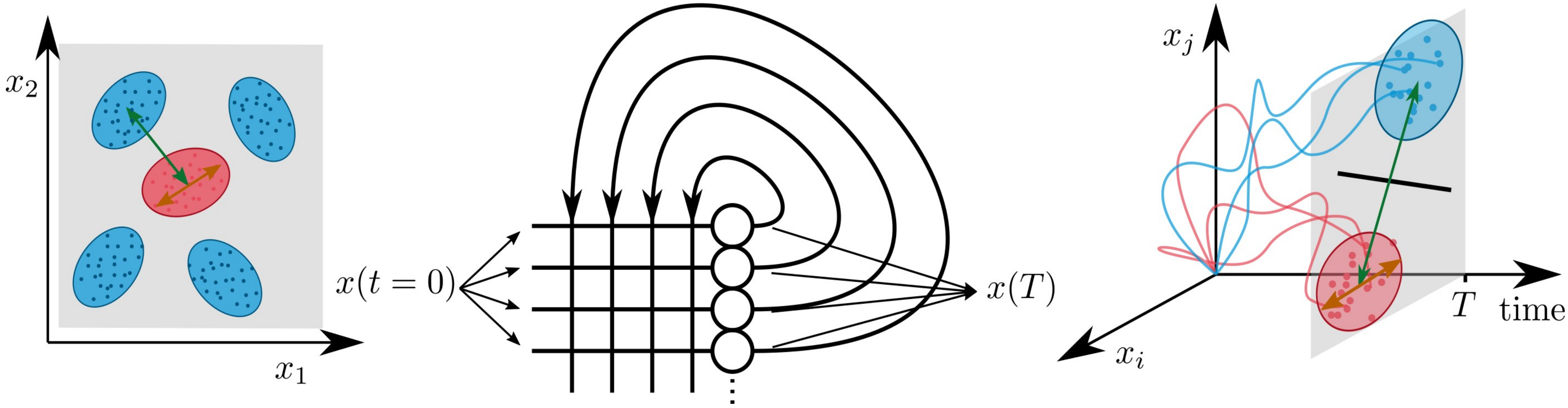
recurrent network





same W
between each pair of time points
= “layers”

COMPARISON OF AI AND BIOLOGICAL NETWORKS



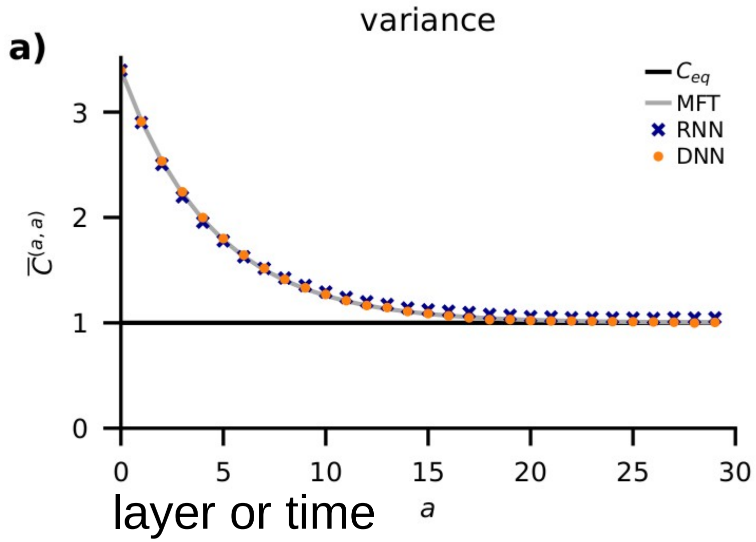
EQUIVALENCE OF DEEP AND RECURRENT NETWORKS

same NNGP, same computational abilities in the limit width N $\rightarrow \infty$

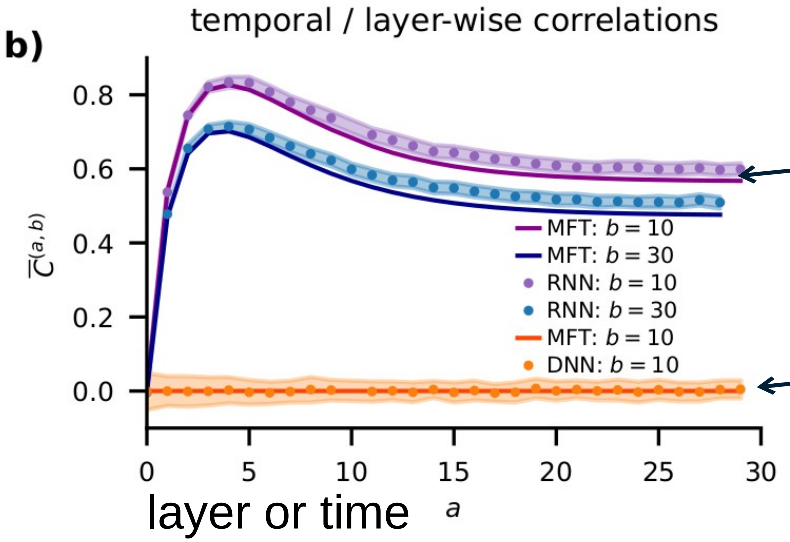
layer index / time index

$$C_{\alpha\alpha}^{(a,b)} = g^2 \left\langle \phi_\alpha^{(a)} \phi_\alpha^{(b)} \right\rangle$$

correlations at same time a=b



correlations at different times a != b



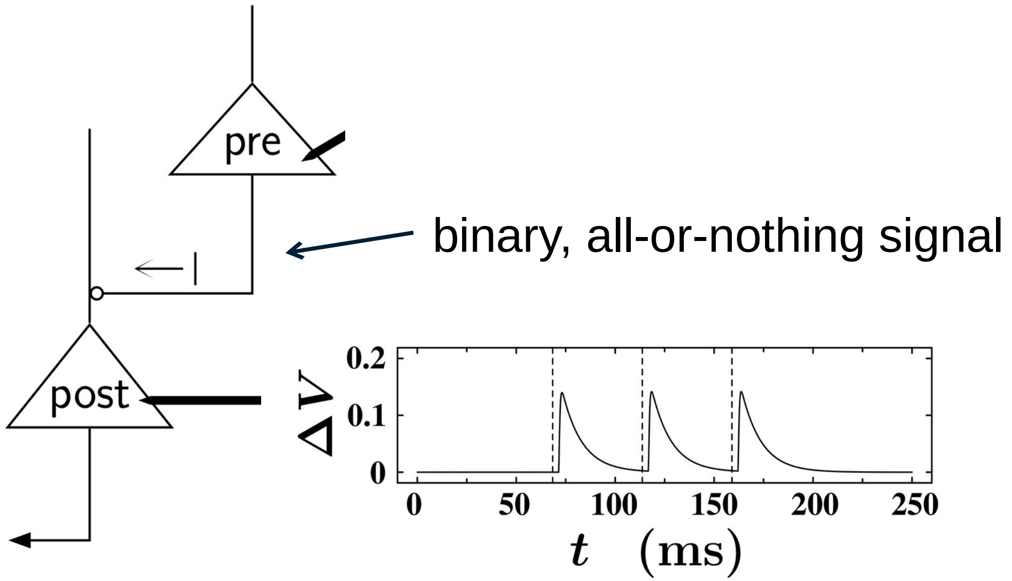
shared weights (recurrent network)

$$W^{(l)} \equiv W$$

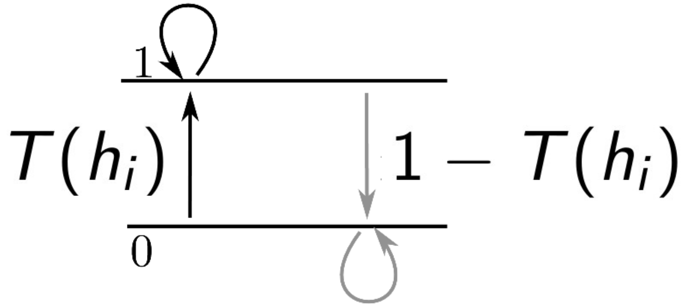
independent weights (deep network) $W^{(l)}$

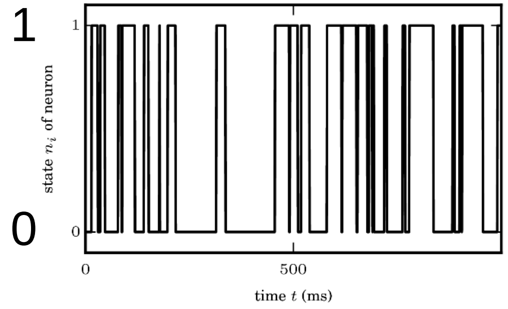
Segadlo et al., Unified field theory for deep and recurrent networks J Stat Mech, 2022

BIOLOGICAL NETWORKS: BINARY INTERACTION



Binary neurons





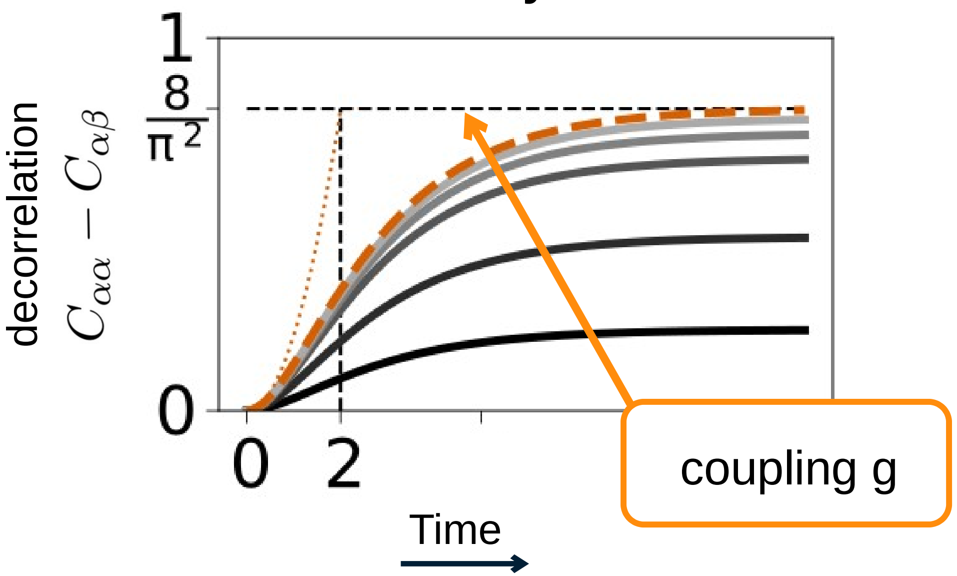
PATTERN SEPARATION

Differences between artificial (continuous rate) and biological (binary) networks

biological recurrent network

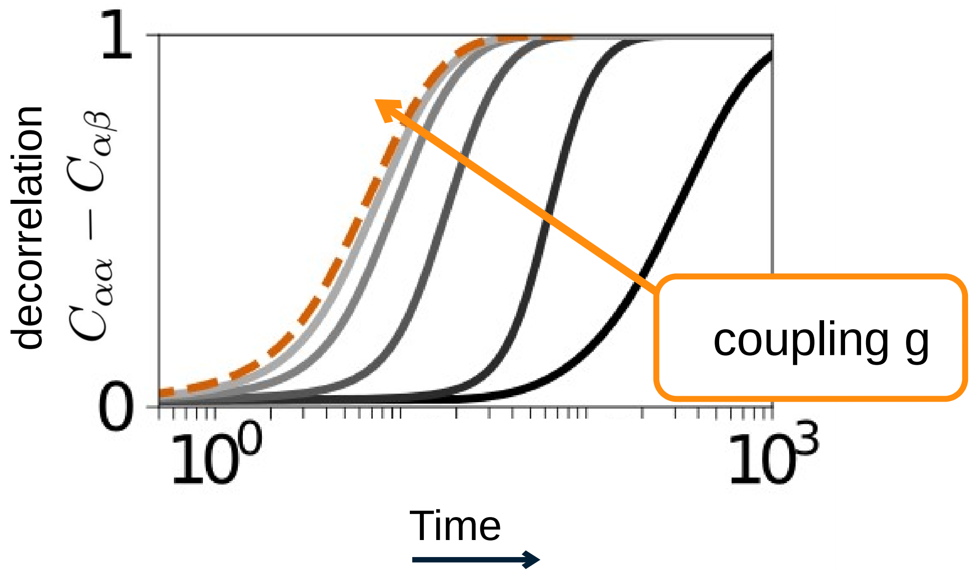
artificial recurrent network

Binary



common mean-field theory for both architectures

Rate

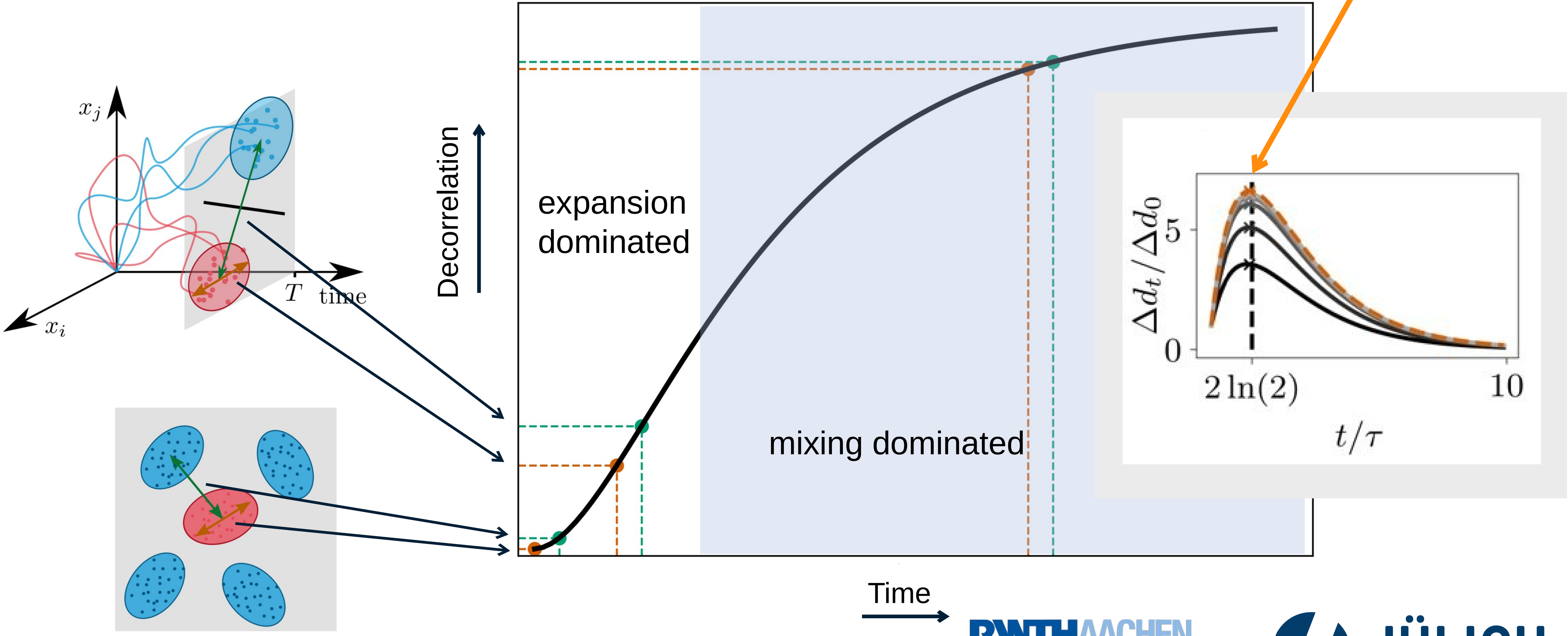


Keup et al., Transient chaotic dimensionality expansion PRX, 2021

PATTERN SEPARATION

Inter-class distance increases compared to intra-class distance

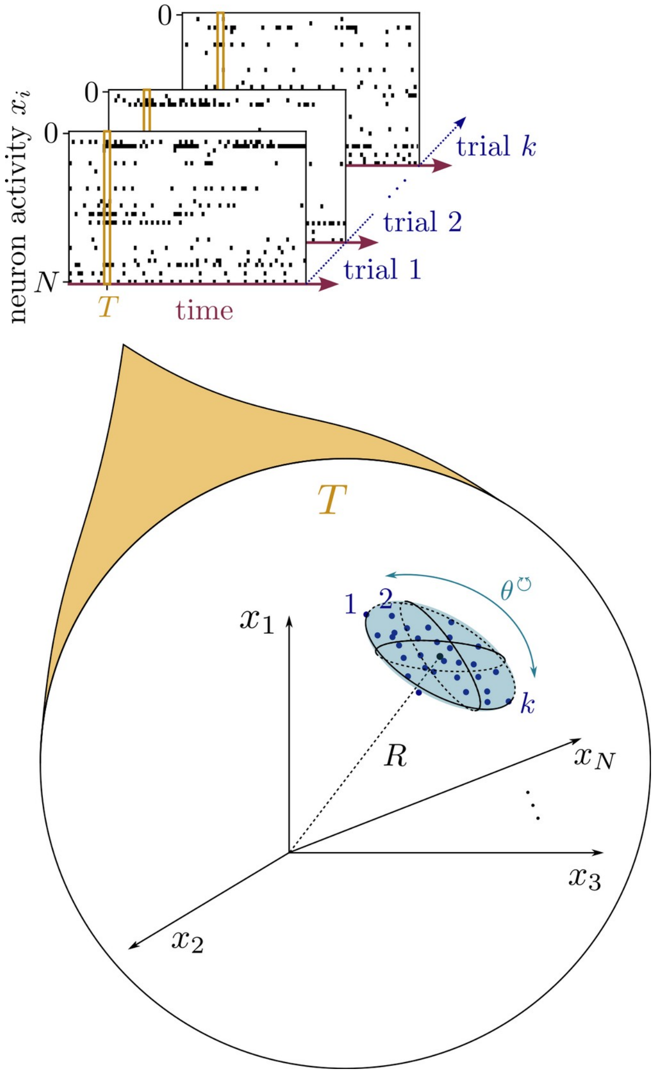
optimal signal after
2 ln(2) ~ 1.5 activations
per neuron



RWTH AACHEN UNIVERSITY

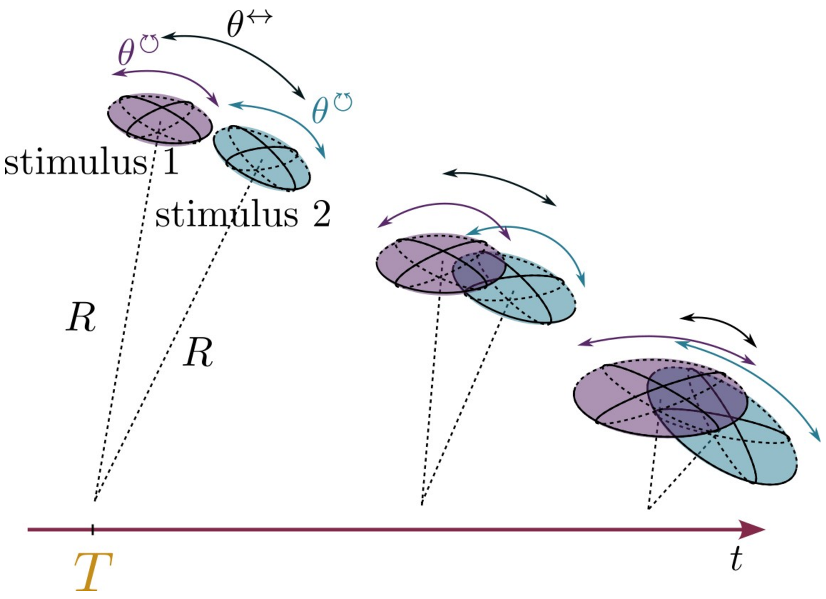
JÜLICH Forschungszentrum

COMPARISON TO THE LIVING (MOUSE) BRAIN

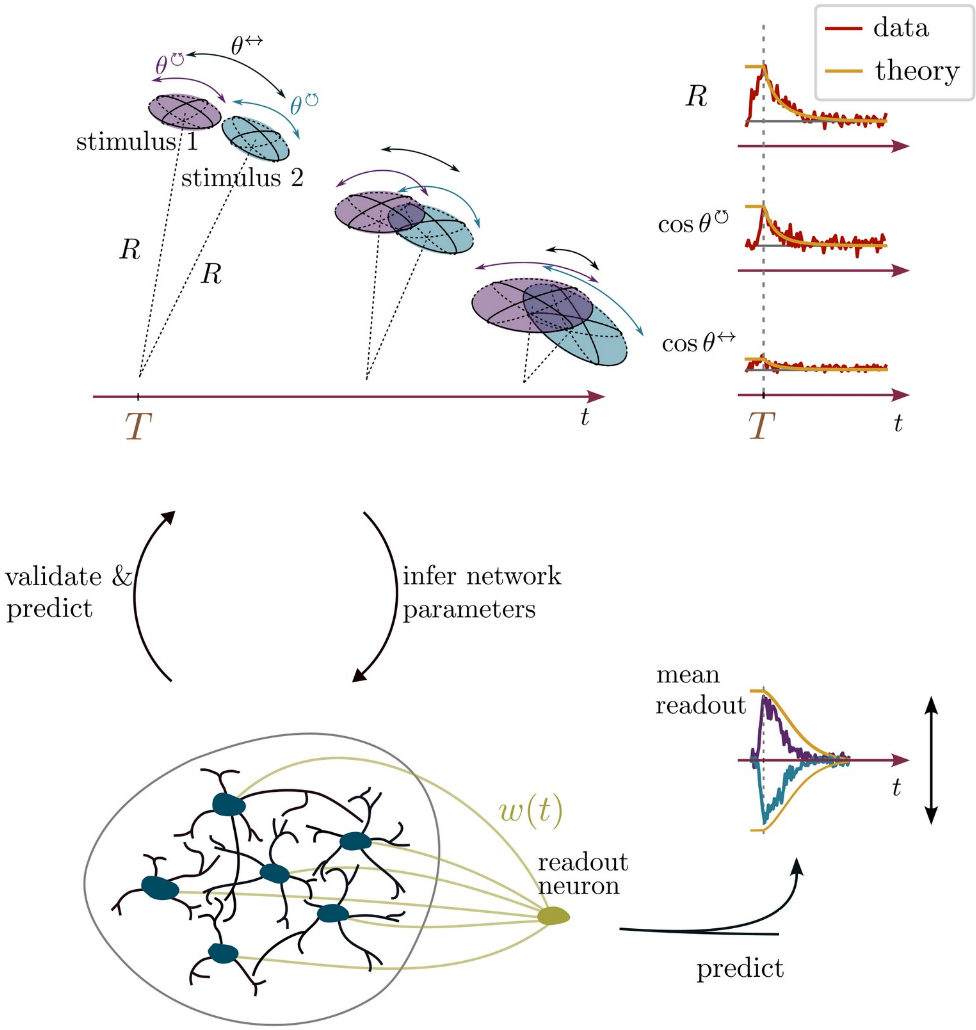


parallel neuropixel recordings in the bahaving mouse
(collaboration Simon Musall, RWTH)

different stimuli
- visual
- tactile



CONSTRAINING A RECURRENT NETWORK MODEL



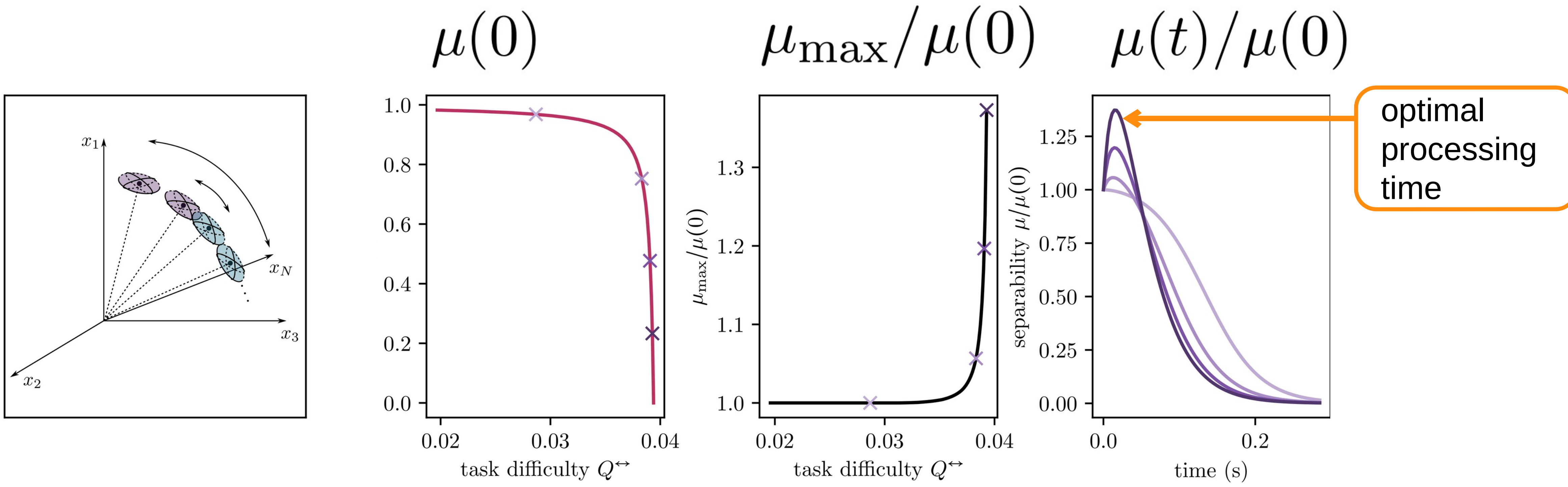
R, both angles ϴ:
- uniquely define parameters of random binary network model

optimally trained readout w
- prediction of separability

→ assess what a downstream neuron may decode

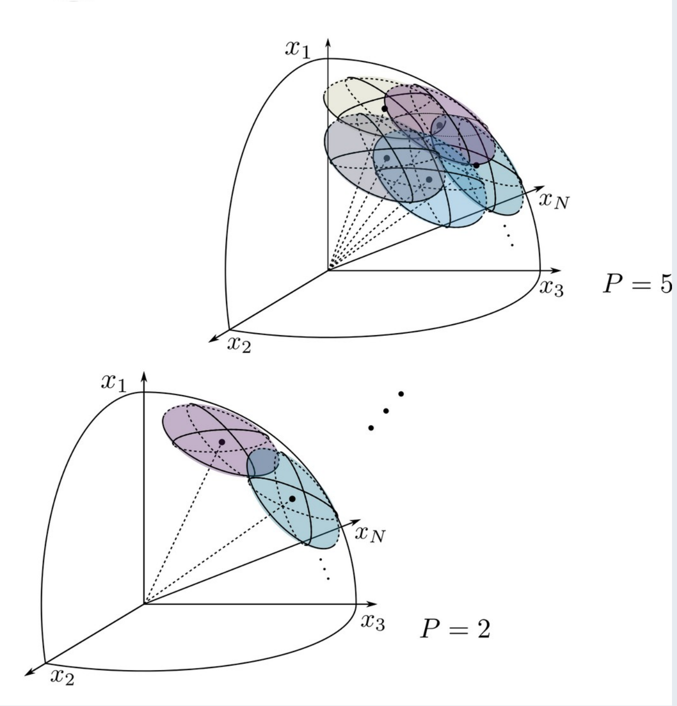
stimulus separability
$\mu(t)$

OPTIMAL PROCESSING TIME FOR HARD TASKS

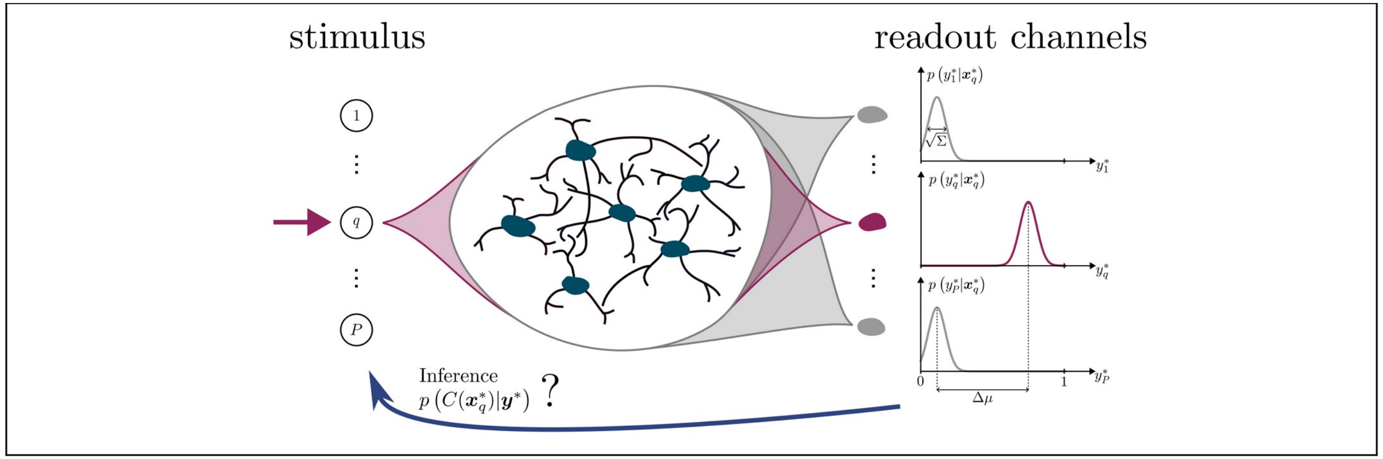


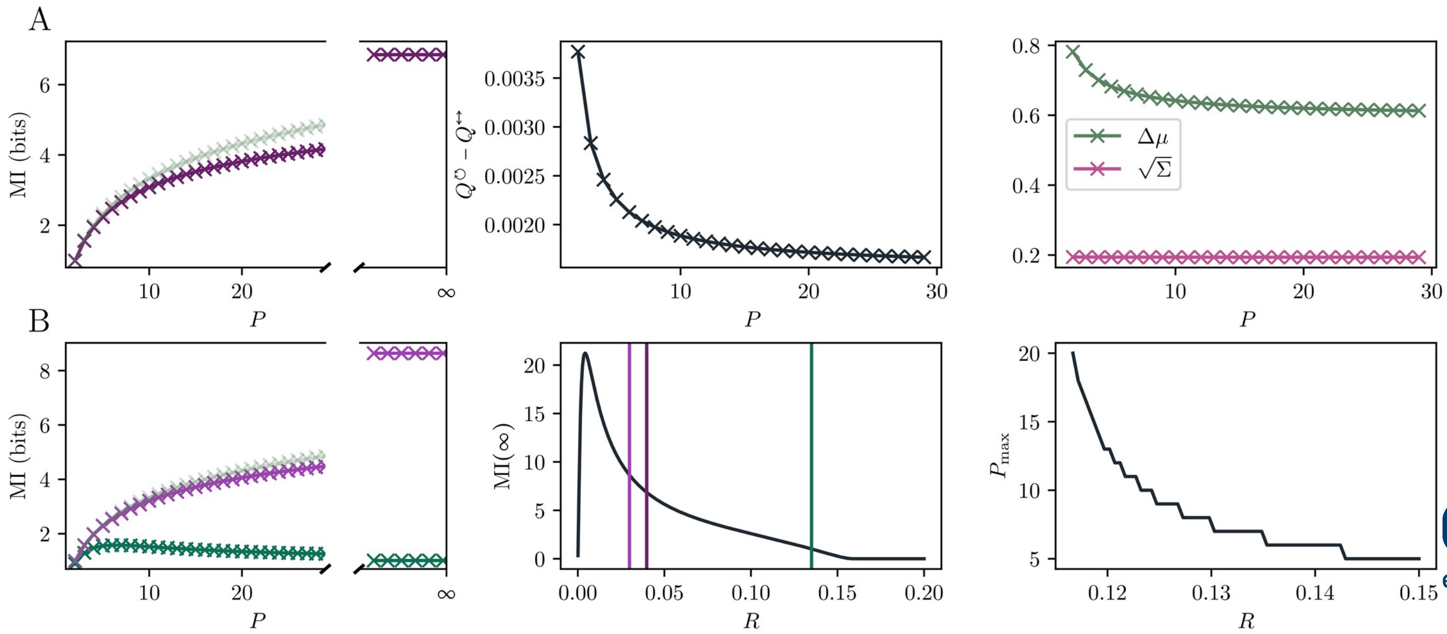
task difficulty
= correlation between classes

ADVANTAGE OF SPARSE CODING: EXTENSIVE INFORMATION GROWTH



crowding of neural space





ACKNOWLEDGMENTS

Dr. Christian Keup

Dr. Tobias Kühn

Javed Lindner

Kirsten Fischer

Fischer, Lindner et al. Critical feature learning in deep networks
ICML 2024 arxiv 2405.10761

Keup, Kühn et al., Transient chaotic dimensionality expansion
PRX, 2021

Segadlo et al., Unified field theory for deep and recurrent networks
J Stat Mech, 2022

Schutzeichel et al. 2024 in prep.

Collaborations:
- Michael Krämer (RWTH)
- Zohar Ringel (HUJI, IL)
- Simon Musall (RWTH)

DFG Deutsche Forschungsgemeinschaft
German Research Foundation

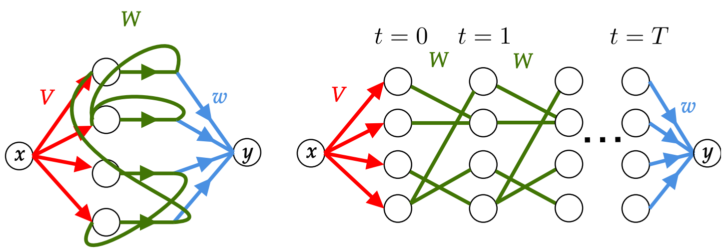
Federal Ministry of Education and Research

RWTH AACHEN UNIVERSITY

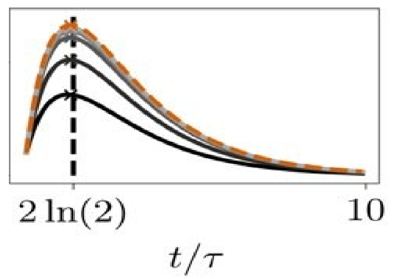
JÜLICH Forschungszentrum

SUMMARY II

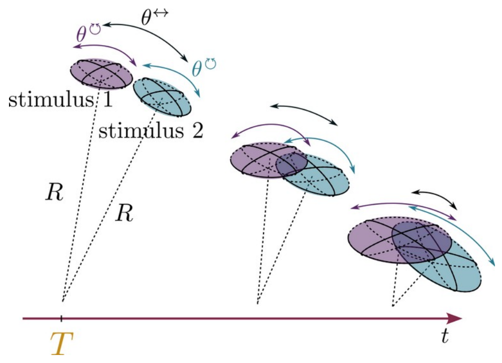
- **deep and recurrent networks:**
  identical large N theory, identical processing capabilities



- **continuous vs discrete (spiking) communication**
  discrete communication results in stereotypical
  optimal processing time



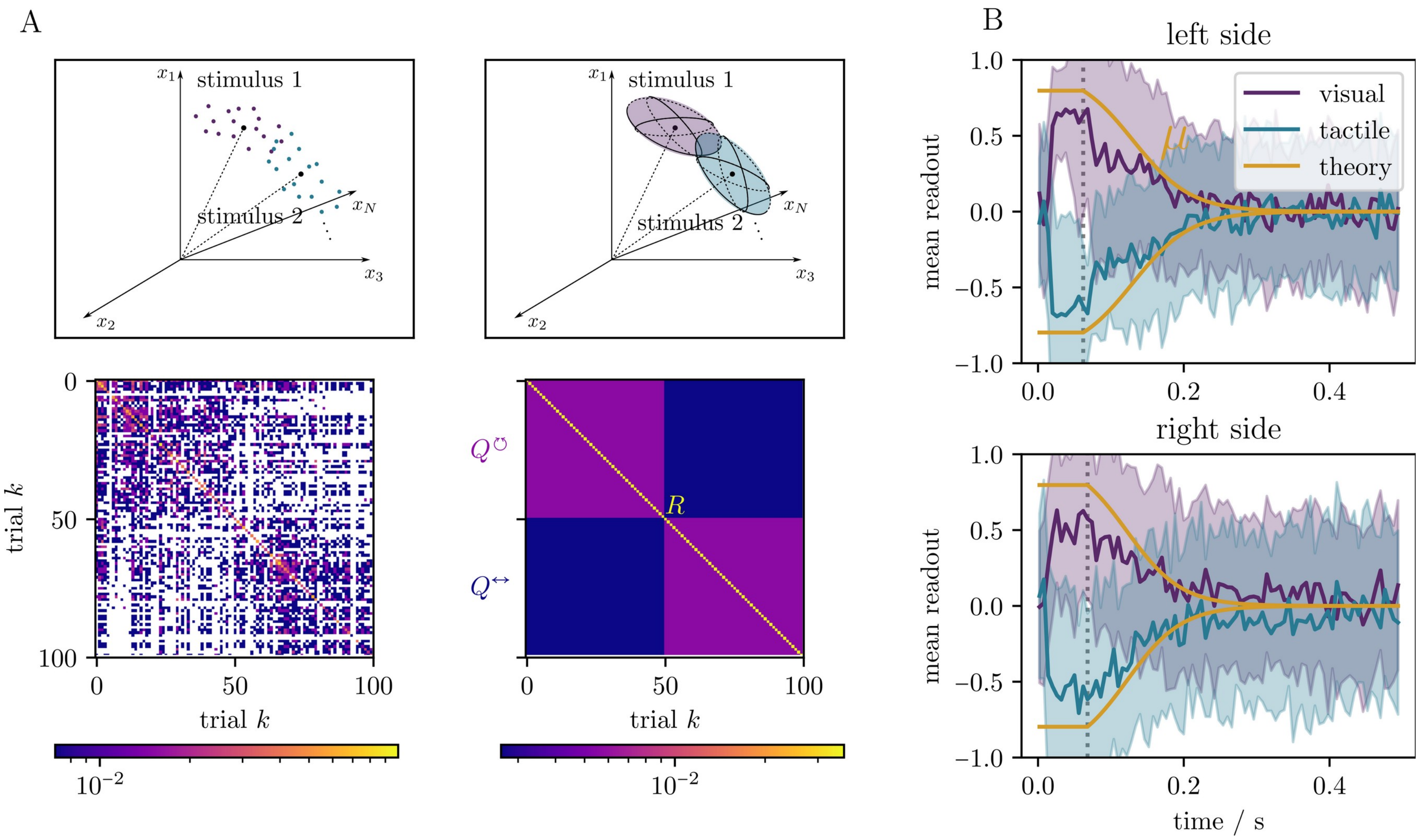
- **signals in the brain**
  optimal transient processing
  nearly extensive information transfer by sparse activity

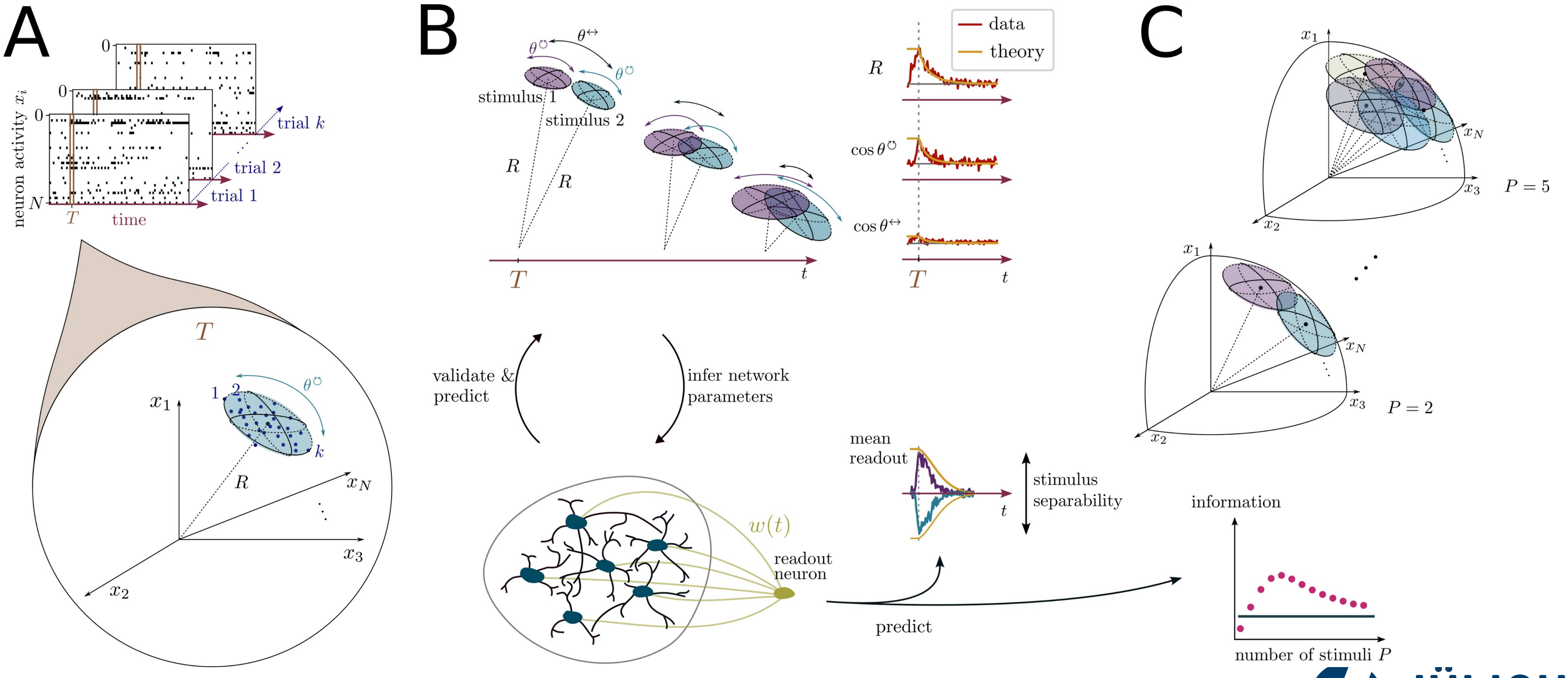


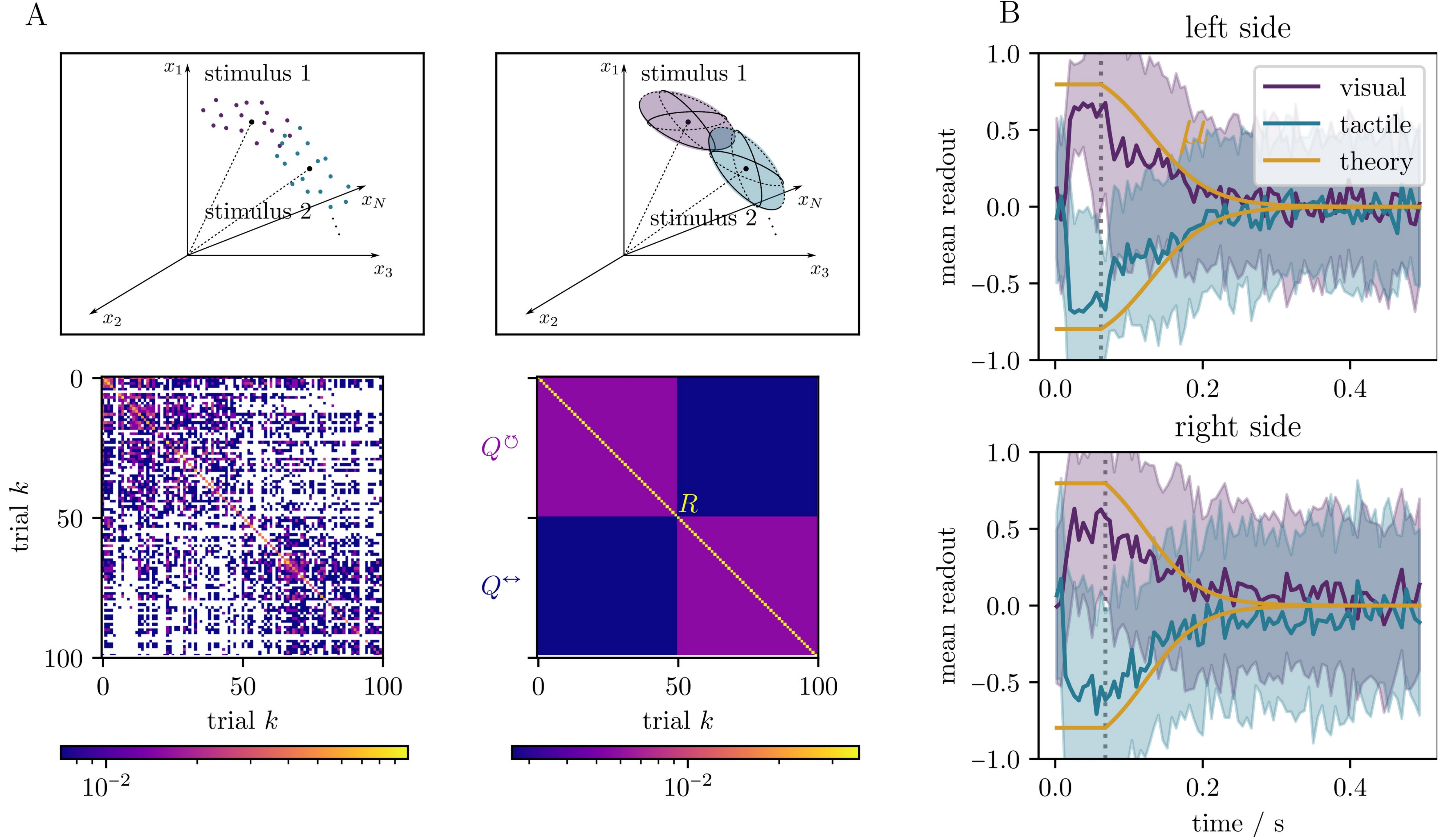
New mailing list on Physics of AI
https://lists.fz-juelich.de/mailman/listinfo/phys4ml

RWTH AACHEN UNIVERSITY

JÜLICH Forschungszentrum


A
stimulus 1
stimulus 2
x_1
x_2
x_3
x_N
trial k
$Q^{\circlearrowright}$
$Q^{\leftrightarrow}$
R
10^{-2}
B
left side
right side
mean readout
time / s
visual
tactile
theory


A
0
0
0
neuron activity x_i
N
T
time
trial k
trial 2
trial 1
T
x_1
1
2
$\theta^{\circlearrowleft}$
k
R
x_N
x_3
x_2
B
$\theta^{\circlearrowleft}$
$\theta^{\leftrightarrow}$
$\theta^{\circlearrowleft}$
stimulus 1
stimulus 2
R
R
T
t
data
theory
R
$\cos\theta^{\circlearrowleft}$
$\cos\theta^{\leftrightarrow}$
T
t
validate &
predict
infer network
parameters
w(t)
readout
neuron
mean
readout
t
stimulus
separability
predict
C
x_1
x_N
x_3
x_2
$P = 5$
x_1
x_N
x_3
x_2
$P = 2$
information
number of stimuli P


A
B
stimulus 1
stimulus 2
x_1
x_2
x_3
x_N
trial k
$Q^{\circlearrowleft}$
$Q^{\leftrightarrow}$
R
10^{-2}
left side
right side
mean readout
time / s
visual
tactile
theory
ω

- learning as Bayesian inference
  - linear regression
  - generalization to networks, selection of networks

- figure for networks (Javed’s poster)

- large N field theory for deep networks

- NNGP theory of networks
  Intuition for the kernel: transformation of similarities
  CIFAR 10 images
  NNGP kernels as function of depths

- feature learning: taking into account data term
- inclusion of data variability (Javed)

- equivalence of deep and recurrent networks
  - intuition
  - justification from large large N limit

- effect of gain function (smooth vs soft) on network properties
- transient dimensionality expansion

OUTLOOK

Community

- collaborations

John Paul Strachan (networking PhD)
Michael Kraemer (physics, RWTH)
Zohar Ringel (physics, Hebrew University)
Alex Alemi (google)

- Phys4ML mailing list

- https://lists.fz-juelich.de/mailman/listinfo/phys4ml
- 150 members

Funding

- past funding

- BMBF project "Renormalized flows" 2020-2023
  2.5 Mio Euro total / ~1 Mio Euro to Juelich / RWTH
- RWTH ERS project (400 kEuro / 1 year)

- application for DFG Research Unit

Lenka Zdeborova (Lausanne)
Bernd Rosenow (Leipzig)
Claudius Gros (Frankfurt)
Caterina De Bacco (Tuebingen)
Peter Sollich (Goettingen)
Michael Kraemer (RWTH)
Zohar Ringel (Hebrew U)

Other activities

- Bocconi University

Marc Mezard builds up computational sciences

- DPG conference 2023

- organized physics meets ML, ~400 attendends

- Special issues

2020 J Phys A Machine learning and statistical physics
https://iopscience.iop.org/article/10.1088/1751-8121/abca75

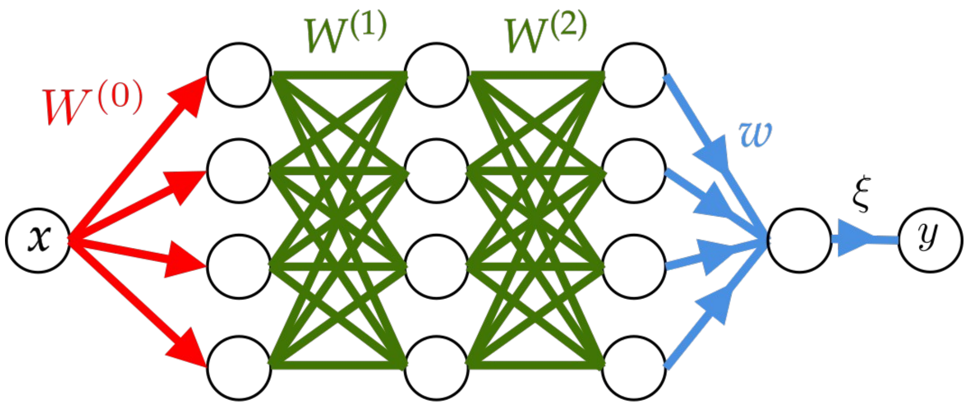
2024 PNAS Machine learning meets physics: A two-way street
https://www.pnas.org/toc/pnas/current

Opportunities for Juelich

- strong in:
  * physics (using methods from there), connection to RWTH
  * computational neuroscience (paradigms of neuronal computation)
  * neuromorphic computing (propose new paradigms)
  * numerical techniques (HPC)

- complimentary to industrial **empirical / applied research** on methods
- requires strong theoreticians and long-term commitment
  to develop coherent theory

THEORY OF LEARNING AND INFERENCE IN DEEP NETWORKS



Goals:
- understand learning and generalization
- prediction of optimal parameters
- implicit bias
- generalization beyond training data-set (transfer learning, important for foundation models)

FEATURE LEARNING THEORY

theory

thermodynamic limit:

$$N \longrightarrow \infty \quad P = \alpha N \to \infty$$

proportional limit

tools:
- large deviation theory
- perturbation expansion $P \ll N$

finite-size:

$$N = \text{finite}$$

$$P = \text{finite}$$

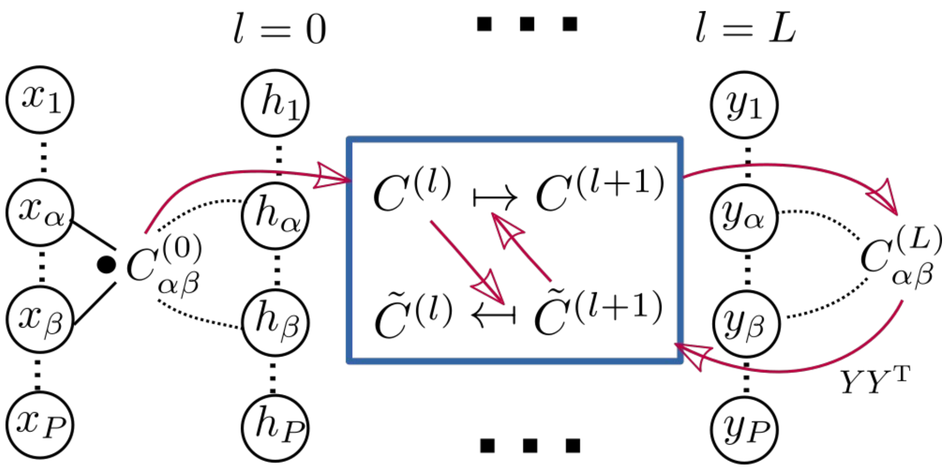
tools:
- field theory
- fluctuation expansion around NNGP

forward mapping

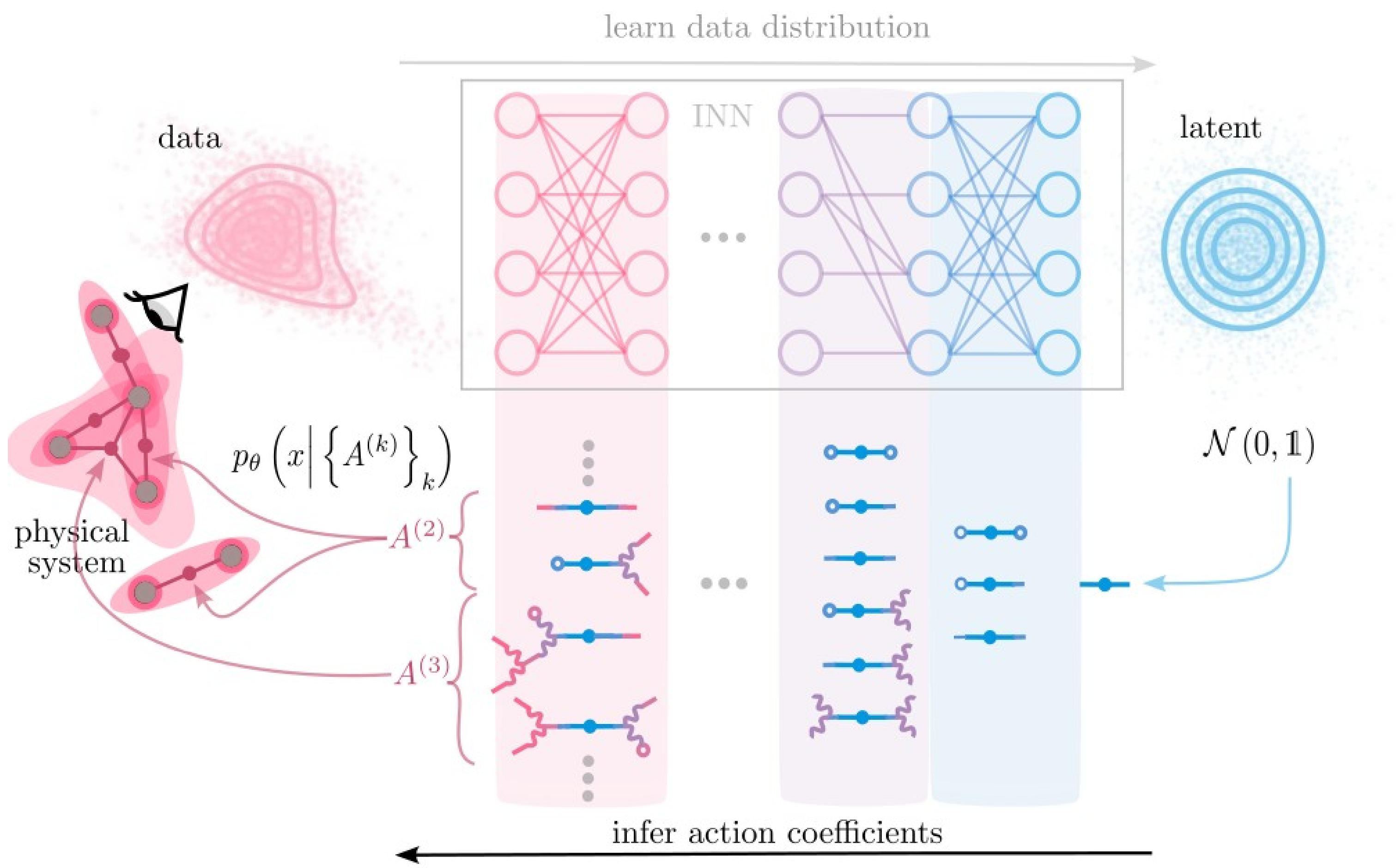
$$C^{(l-1)} \mapsto C^{(l)}$$

backward mapping

$$\tilde{C}^{(l+1)} \mapsto \tilde{C}^{(l)}$$



EXPLAINABLE AI



- use of INN for unsupervised learning of data statistics

- extraction of theory: build up of interactions in hierarchical manner

Merger et al., Learning interacting theories from data, Phys Rev X, 2023
press release, radio interview

LARGE-N FIELD THEORY

- joint distribution of outputs for training set

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \left\langle \prod_{\alpha=1}^{D} \delta[y_\alpha - \Psi(\theta, x_\alpha)] \right\rangle_{\theta \sim N}$$

- disorder average, auxiliary variable C enforced by $\tilde{C}$

$$\boldsymbol{Y} \,|\, \boldsymbol{X} \quad \sim \quad \langle N(0, C) \rangle_{C, \tilde{C}}$$

- distribution of $(C, \tilde{C}) \sim \exp\left(S(C, \tilde{C})\right)$

large N limit

$$S(C, \tilde{C}) = \boxed{N} \left[-\tilde{C}^{\mathrm{T}} C + \Omega(\tilde{C} \,|\, C) \right]$$

$$\begin{aligned}\Omega(\tilde{C} \,|\, C) \quad &= \quad \sum_{l=1}^{L+1} \ln \left\langle e^{\tilde{C}^{(l)}_{\alpha\beta} g^2 \, \phi^{(l-1)}_\alpha \, \phi^{(l-1)}_\beta} \right\rangle_{h^{(l-1)} \sim N(0, C^{(l-1)})} \\ &+ \quad \tilde{C}^{(0)}_{\alpha\beta} \, \frac{g_0^2}{N_{\text{in}}} \sum_{i=1}^{N_{\text{in}}} x_{i,\alpha} \, x_{i,\beta}\end{aligned}$$

Slides of oxford talk follow here

LEARNING IN NEURAL NETWORKS

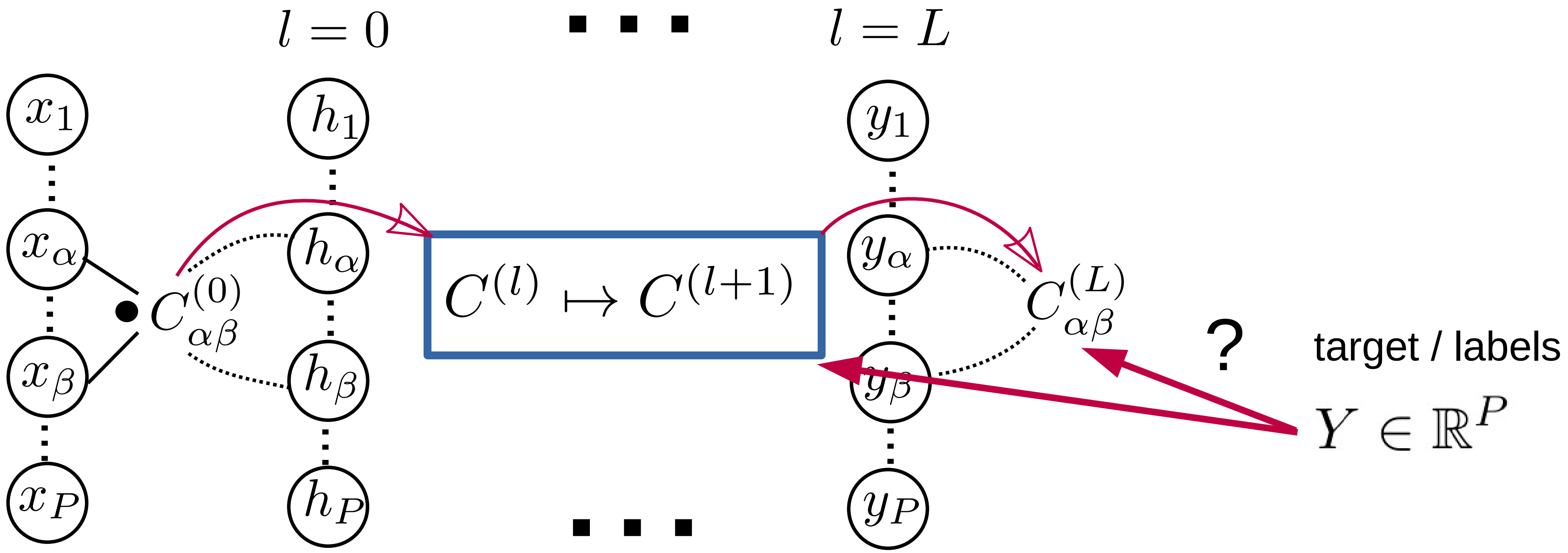
Motivation

There are two regimes in the theory of neural networks:

- lazy learning (Chizat et al., 2019)
  → neural network Gaussian process (NNGP) (Neal 1994; Williams 1998; Lee et al., 2018)
    - equivalent to random feature regression (Mei et al., 2022)
  → neural tangent kernel (NTK) (Jacot et al., 2018)
    - equivalent to linearization in weights

- **feature learning**
  → network parameters adapt to task and network learns features of task
  → networks typically show better performance (Geiger et al., 2020)
  - related works:
    Naveh & Ringel 2021; Zavatone-Veth, …, Pehlevan (2021);
    Li & Sompolinsky, 2021; Hanin & Zlokapa, 2023; Seroussi et al., 2023;
    Pacelli et al., 2023; Cui et al., 2023

HOW TO GO FROM NNGP TO FEATURE LEARNING?



1. Recover NNGP for width N $\rightarrow \infty$, P = const. from a large deviation principle

2. Minimal extension of this approach to the proportional limit N, P $\rightarrow \infty$ to obtain feature learning

3. Expose relation to networks at finite N: optimal adaptation

INTERMEDIATE KERNELS: NATURAL ORDER PARAMETERS

$$\prod_{\alpha i} \delta[-h_{\alpha i} + \sum_j W_{ij}\phi_{\alpha j}] = \prod_{\alpha i} \int_{-i\infty}^{i\infty} \frac{d\tilde{h}_{\alpha i}}{2\pi i} \exp\left(\tilde{h}_{\alpha i}[-h_{\alpha i} + \sum_j W_{ij}\phi_{\alpha j}]\right)$$

neurons decouple, quadratic

$$\left\langle \exp\left(\sum_{\alpha i} \tilde{h}^{(l)}_{\alpha i} W^{(l)}_{ij} \phi^{(l-1)}_{\alpha j}\right) \right\rangle_{W^{(l)}} = \exp\left(\frac{1}{2}\sum_{\alpha\beta} \boxed{\sum_i \tilde{h}^{(l)}_{\alpha i}\tilde{h}^{(l)}_{\beta i}} \frac{g_l}{N}\sum_j \phi^{(l-1)}_{\alpha j}\phi^{(l-1)}_{\beta j}\right)$$

$$C^{(l)}_{\alpha\beta} = \frac{g_l}{N}\phi^{(l-1)}_\alpha \cdot \phi^{(l-1)}_\beta$$

$$h^{(l)}_{\alpha i} | C^{(l)} \overset{\text{i.i.d. in } i}{\sim} \mathcal{N}(0, C^{(l)})$$

suggest **concentration** of auxiliary variables C for large N
given C: neurons **decouple**, preactivations i.i.d. Gaussian

qualitatively similar approaches: Sompolinsky & Zippelius (1982) (spin glasses)
Schuecker et al.. (2016, 2018), Crisanti et al. (2018) (cont.-time RNNs)

LARGE DEVIATION APPROACH

scaling form of cumulant-generating function, limit exists

independent of N

$$\lim_{N\to\infty} \boxed{N^{-1}}\mathcal{W}(\boxed{N}K|C^{(l-1)}) = \lambda(K|C^{(l-1)}) \equiv \ln\left\langle \exp\left(g_l \phi^{(l-1)\mathsf{T}} K \phi^{(l-1)}\right)\right\rangle_{\mathcal{N}(0,C^{(l-1)})}$$

Gärtner-Ellis theorem (e.g., Touchette 2009)

$$\begin{aligned} -\ln p\left(C^{(l)}|C^{(l-1)}\right) &\simeq \sup_K \boxed{N}\left[\mathrm{tr}K^{\mathsf{T}}C^{(l)} - \lambda(K|C^{(l-1)})\right] \\ &=: \Gamma\left(C^{(l)}|C^{(l-1)}\right) \end{aligned}$$

rate function

supremum condition: **forward propagation of kernel** $C^{(l-1)} \mapsto C^{(l)}$

$$\boxed{\begin{aligned} C^{(l)} &= \lambda'\left(K^{(l)}|C^{(l-1)}\right) \\ &= g_l\left\langle \phi^{(l-1)}\phi^{(l-1)\mathsf{T}}\right\rangle_{\mathcal{P}} \end{aligned}}$$

non-Gaussian measure

$$\langle \ldots \rangle_{\mathcal{P}} \propto \left\langle \ldots \exp\left(g_l \boxed{\phi^{(l-1)\mathsf{T}} K^{(l)} \phi^{(l-1)}}\right)\right\rangle_{\mathcal{N}(0,C^{(l-1)})}$$

MAXIMUM A POSTERIORI (MAP) ESTIMATE OF KERNELS

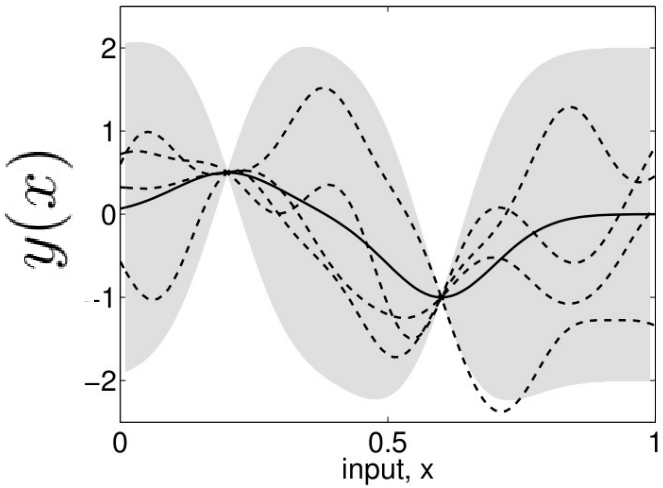
Bayes

$$p(C|Y) \propto p(Y,C) \equiv \mathcal{N}\left(Y|0, C^{(L)} + \kappa\mathbb{I}\right) p(C)$$

data term $\propto P$ net prior $\propto N$

$$\mathcal{S}(C) := \ln p(C|Y) \overset{\text{l.d.p.}}{\simeq} \mathcal{S}_{\mathrm{D}}(C^{(L)}) - \sum_{l=1}^{L} \Gamma(C^{(l)}|C^{(l-1)})$$

$$\mathcal{S}_{\mathrm{D}}(C^{(L)}) := \ln \mathcal{N}(Y|0, C^{(L)} + \kappa\mathbb{I})$$



from Rasmussen & Williams (2006)

MAP ESTIMATE OF KERNELS

Recovering the NNGP

$$\left.\begin{array}{ll} N & \longrightarrow \infty \\ P & \text{finite} \end{array}\right\}$$

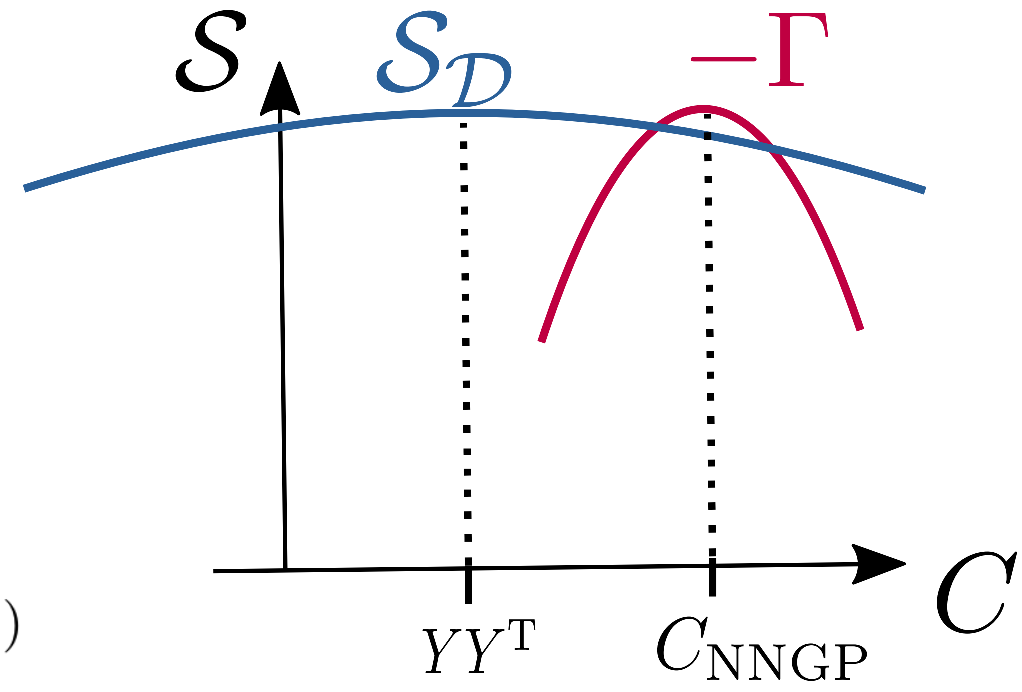
data term $\propto P$ net prior $\propto N$

$$\mathcal{S}(C) := \ln p(C|Y) \overset{\text{l.d.p.}}{\simeq} \mathcal{S}_{\mathrm{D}}(C^{(L)}) - \sum_{l=1}^{L} \Gamma(C^{(l)}|C^{(l-1)})$$

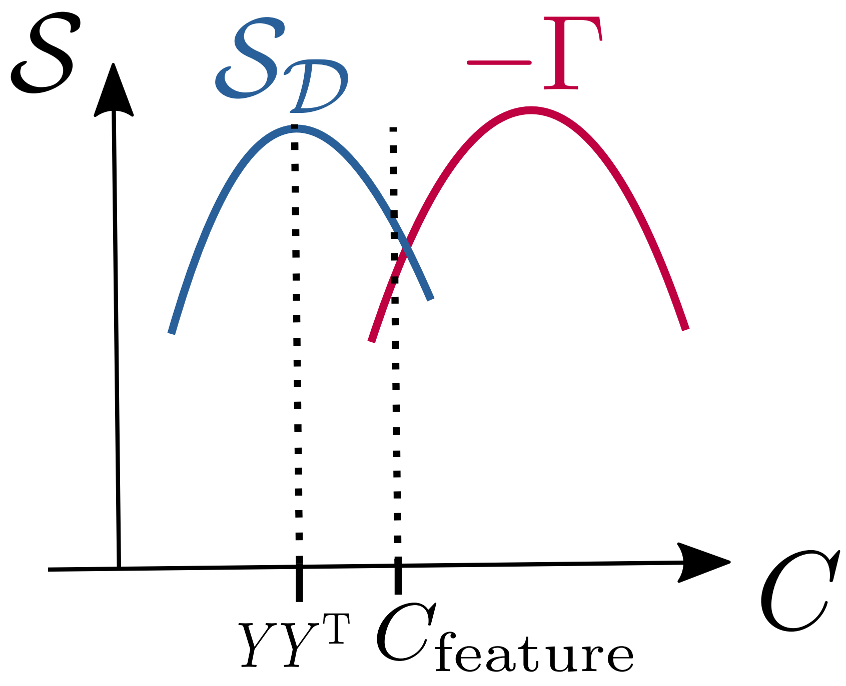
$$0 \overset{!}{=} \mathcal{S}'(C) \simeq -\Gamma'(C^{(l)}|C^{(l-1)}) \equiv -\tilde{C}^{(l)} \quad \forall l$$

$$\langle \ldots \rangle_{\mathcal{P}} \propto \langle \ldots \exp(\ldots \tilde{C} \ldots) \rangle = \langle \ldots \rangle_{\mathcal{N}}$$

from sup condition: $$C^{(l)} = g_l \left\langle \phi^{(l-1)\mathsf{T}} \phi^{(l-1)\mathsf{T}} \right\rangle_{\mathcal{N}(0, C^{(l-1)})}$$ **NNGP**



MAP ESTIMATE OF KERNELS



Proportional limit -> feature learning

$$\left.\begin{array}{l} N \longrightarrow \infty \\ P = \alpha N \rightarrow \infty \end{array}\right\}$$

$$\boxed{C^{*(l)}: \quad \frac{\partial \mathcal{S}}{\partial C^{(l)}} \overset{!}{=} 0}$$

$$\mathcal{S}(C) := \ln p(C|Y) \overset{\text{l.d.p.}}{\simeq} \mathcal{S}_{\mathrm{D}}(C^{(L)}) - \sum_{l=1}^{L} \Gamma(C^{(l)}|C^{(l-1)})$$

output discrepancy: $l = L$

$$\tilde{C}^{(L)} = \mathcal{S}'_{\mathrm{D}}(C^{(L)}) \overset{\text{MAP}}{\simeq} \frac{1}{2\kappa^2} \langle (y - h^{(L)})\,(y - h^{(L)})^{\mathsf{T}} \rangle - \frac{1}{2\kappa}\mathbb{I}$$

"error signal"

back propagation: $1 \leq l < L$

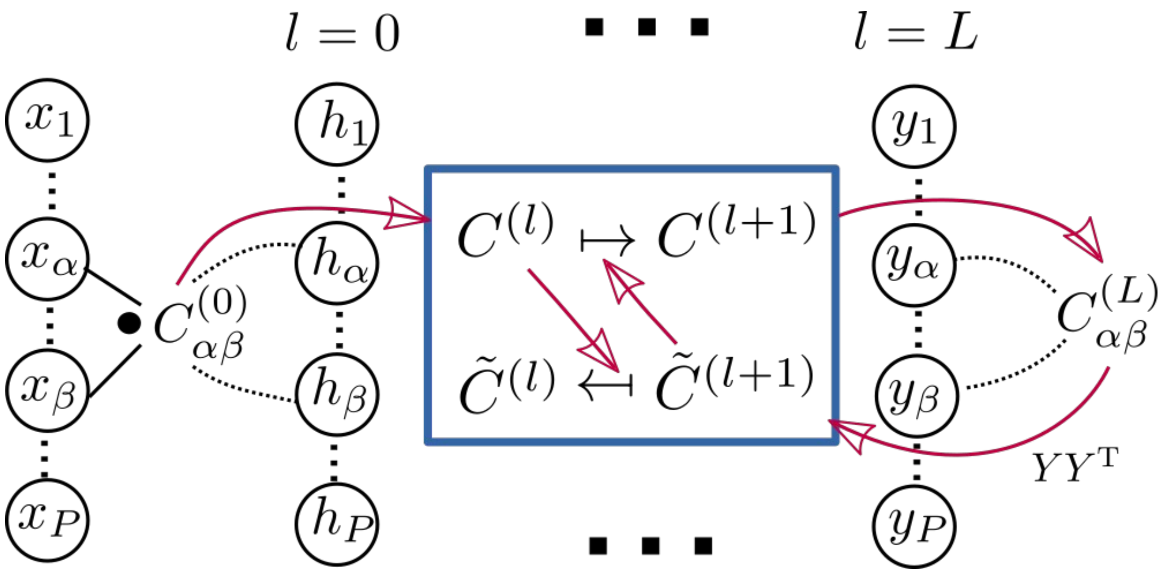
$$\tilde{C}^{(l+1)} \mapsto \tilde{C}^{(l)}$$

$$0 \overset{!}{=} \frac{\partial \mathcal{S}(C)}{\partial C^{(1 \leq l < L)}} = -\underbrace{\Gamma'(C^{(l)}|C^{(l-1)})}_{\tilde{C}^{(l)}} - \underbrace{\frac{\partial \Gamma(C^{(l+1)}|C^{(l)})}{\partial C^{(l)}}}_{\simeq -\frac{\partial}{\partial C^{(l)}} \ln p(C^{(l+1)}|C^{(l)})}$$

PAIR OF FORWARD-BACKWARD EQUATIONS

initial cond.

$$C^{(0)} = \frac{g_0}{D} X X^\mathsf{T}$$



final cond.

$$\tilde{C}^{(L)} = \frac{\partial}{\partial C^{(L)}} \ln \mathcal{N}(y|0, C^{(L)} + \kappa \mathbb{I})$$

forward equation

$$C^{(l+1)} = g_l \left\langle \phi^{(l)} \phi^{(l)\mathsf{T}} \right\rangle_{\mathcal{P}_l}$$

$$\langle \ldots \rangle_{\mathcal{P}_l} \propto \left\langle \ldots \exp\left(\frac{g_{l+1}}{N} \phi^{(l)\mathsf{T}} \tilde{C}^{(l+1)} \phi^{(l)}\right) \right\rangle_{\mathcal{N}(0, C^{(l)})}$$

backward equation

$$\tilde{C}^{(l)} = \frac{\partial}{\partial C^{(l)}} \ln p(C^{(l+1)} | C^{(l)})$$

similar structure as in Seroussi & Ringel 2023 Nat. Comm.

SPECIAL CASE: DEEP LINEAR NETWORK

$$h_\alpha^{(L)} = \{ \prod_{l=0}^{L} W^{(l)} \} \, x_\alpha$$

$$\Gamma(C^{(l)} | C^{(l-1)}) = \mathrm{KL}(\mathcal{N}(0, C^{(l)}) \,||\, \mathcal{N}(0, g_l C^{(l-1)}))$$

consistent with Yang, …, Aitchison (2023)

forward mapping

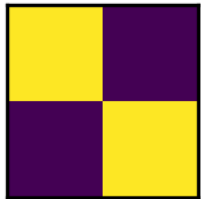
$C^{(l-1)} \mapsto C^{(l)}$ $\qquad C^{(0)} = \frac{g_0}{D} X X^\mathsf{T}$ $\qquad C^{(l)} = C^{(l-1)} \boxed{T^{(l)}}$

backward mapping

$\tilde{C}^{(l+1)} \mapsto \tilde{C}^{(l)}$ $\qquad \tilde{C}^{(L)} = \mathcal{S}'_\mathrm{D}(C^{(L)})$ $\qquad \tilde{C}^{(l-1)} = \boxed{T^{(l)}} \tilde{C}^{(l)}$

$$T^{(l)} = g_l \left[\mathbb{I} - 2 \frac{g_l}{N} \tilde{C}^{(l)} C^{(l-1)} \right]^{-1}$$

rank one correction towards target



$$C_{\alpha\beta}^{(L)} \simeq g_L \, C_{\alpha\beta}^{(L-1)} + \frac{1}{N} \left(\boxed{Y Y^\mathsf{T}} - C^{(L)} \right)_{\alpha\beta}$$

consistent with Li & Sompolinsky (2021)

GENERAL CASE: NON-LINEAR DEEP NETWORK

perturbative treatment

$$\langle \ldots \rangle_{\mathcal{P}} \propto \left\langle \ldots \exp\left(\frac{g}{N}\,\phi^{\mathsf{T}}\tilde{C}\phi\right)\right\rangle_{\mathcal{N}(0,C)}$$

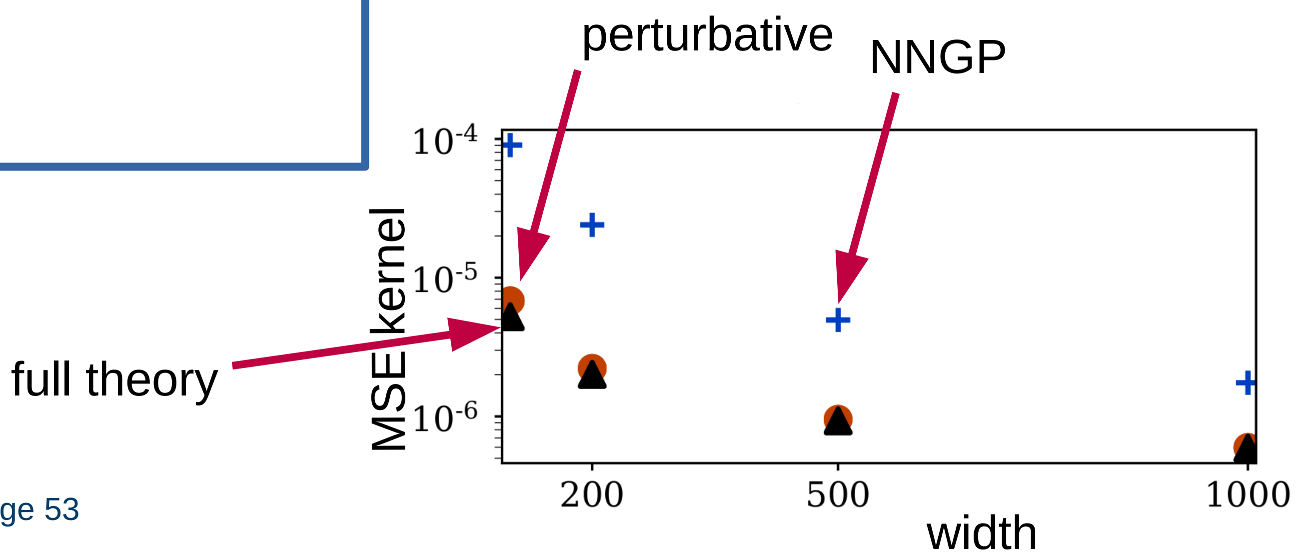
$$\simeq \left\langle \ldots \left[1+\frac{g}{N}\,\phi^{\mathsf{T}}\tilde{C}\phi\right]+\mathcal{O}(N^{-2})\right\rangle_{\mathcal{N}(0,C)}$$

$$C_{\alpha\beta}^{(l+1)} = g_{l+1}\left\langle \phi_{\alpha}^{(l)}\phi_{\beta}^{(l)}\right\rangle_{\mathcal{N}(0,C^{(l)})} + \frac{g_{l+1}^{2}}{N}\sum_{\gamma,\delta} V_{\alpha\beta,\gamma\delta}^{(l)}\,\tilde{C}_{\gamma\delta}^{(l+1)} + \mathcal{O}\left(N^{-2}\right)$$

$$V_{\alpha\beta,\gamma\delta}^{(l)} := \langle \phi_{\alpha}^{(l)}\phi_{\beta}^{(l)},\phi_{\gamma}^{(l)}\phi_{\delta}^{(l)}\rangle_{\mathcal{N}(0,C^{(l)})}^{c}$$

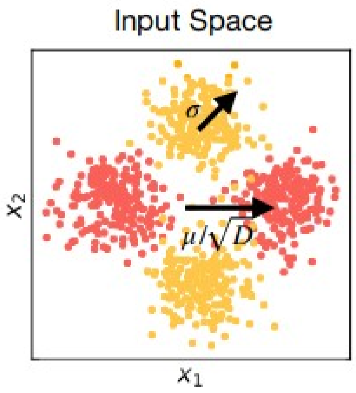
$$\tilde{C}_{\alpha\beta}^{(l)} = \mathcal{G}_{\alpha\beta}\,\tilde{C}_{\alpha\beta}^{(l+1)} + \delta_{\alpha\beta}\sum_{\gamma}\mathcal{H}_{\gamma\alpha}\tilde{C}_{\gamma\alpha}^{(l+1)} + \mathcal{O}(\tilde{C}^{2})$$

check against numerics (linear case)



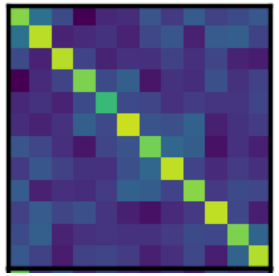
KERNEL ADAPTATION FOR XOR TASK

Numerical evaluation



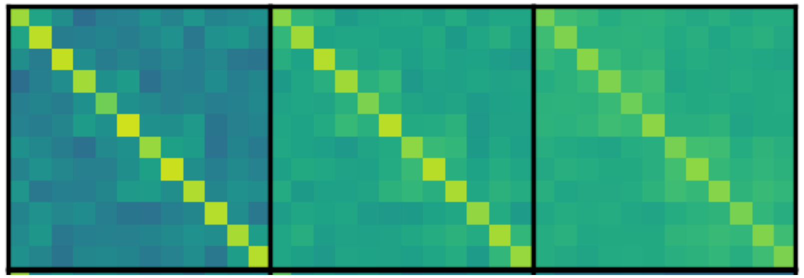
Refinetti et al., ICML 2021

Input $C^{(0)}$



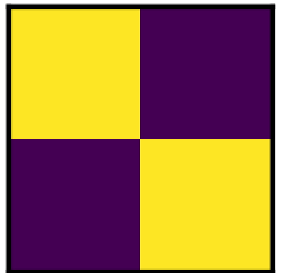
layers

Feature-corrected kernels

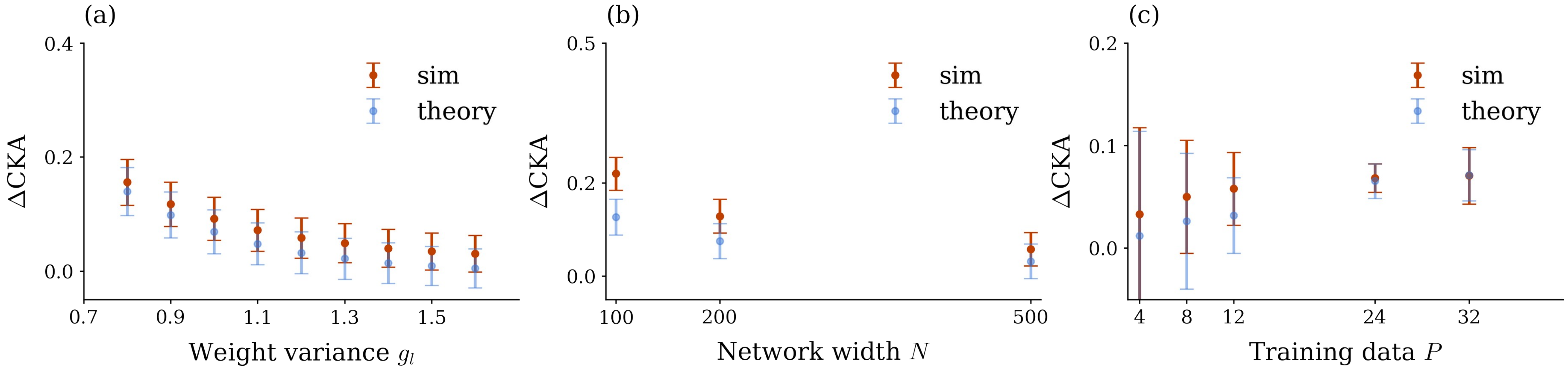


CKA = cosine similarity

Target



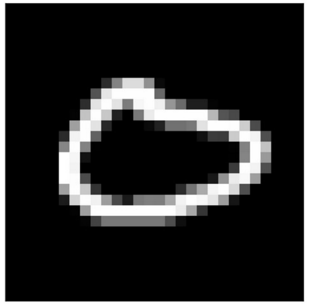
$YY^{\top}$



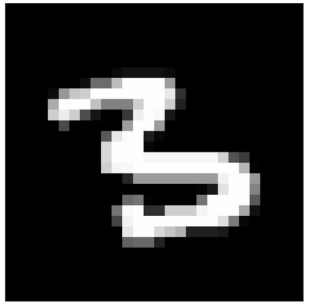
MNIST – CLASSIFICATION BETWEEN 0'S AND 3'S

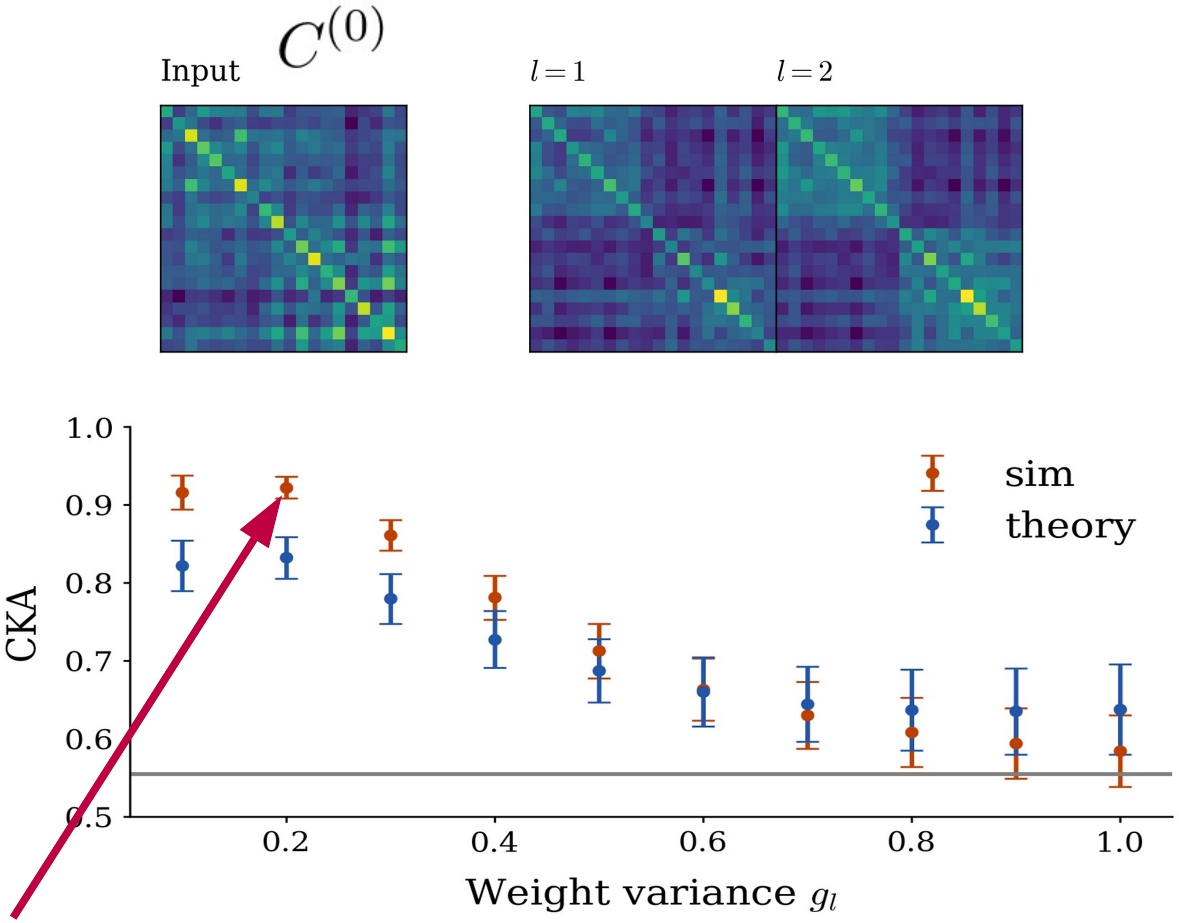
Numerical evaluation

there is no constraint on the input data



vs





optimal alignment

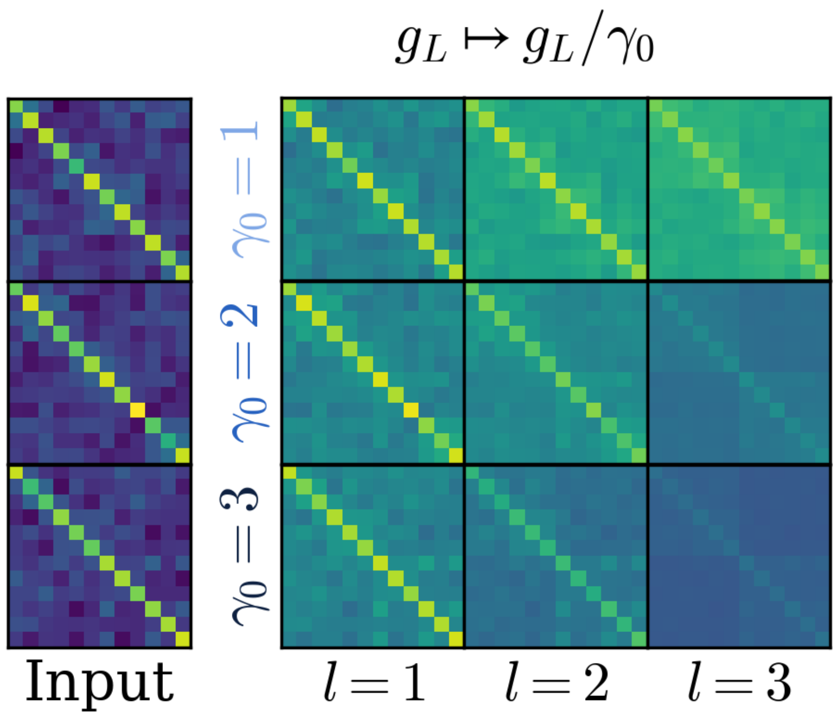
OUTPUT SCALING ENHANCES FEATURE LEARNING

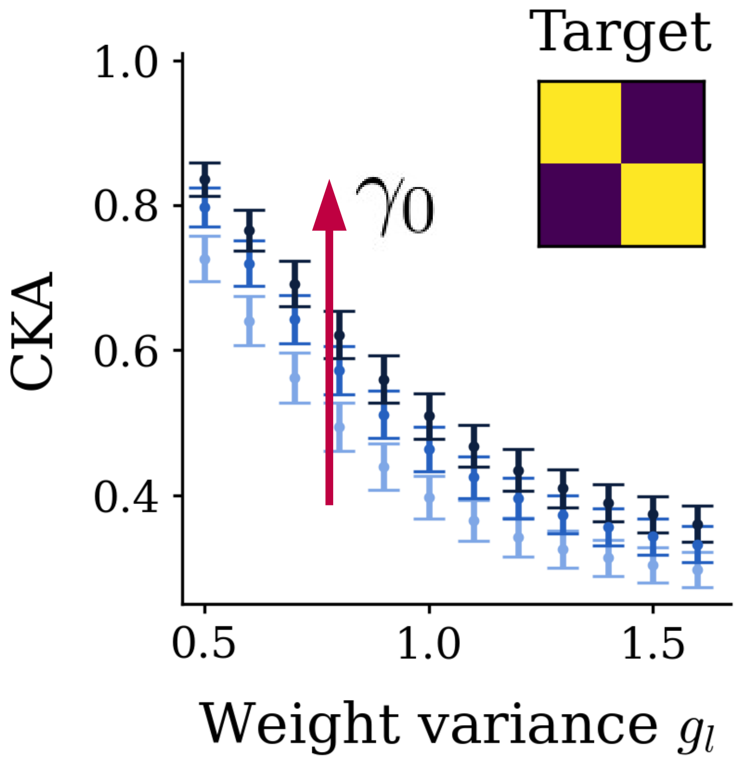
Downscaling of output layer increases corrections of kernels

dependence on output scale

$$\tilde{C}^{(L)} \overset{\text{MAP}}{\simeq} \frac{1}{2\kappa^2} \langle (y - h_\alpha^{(L)}/\sqrt{\gamma_0})\,(y - h^{(L)}/\sqrt{\gamma_0})^\mathsf{T} \rangle - \frac{1}{2\kappa}\mathbb{I}$$

$$\overset{\gamma_0 \to \infty}{\to} \frac{1}{2\kappa^2}\left(yy^\mathsf{T} - \kappa\mathbb{I}\right)$$

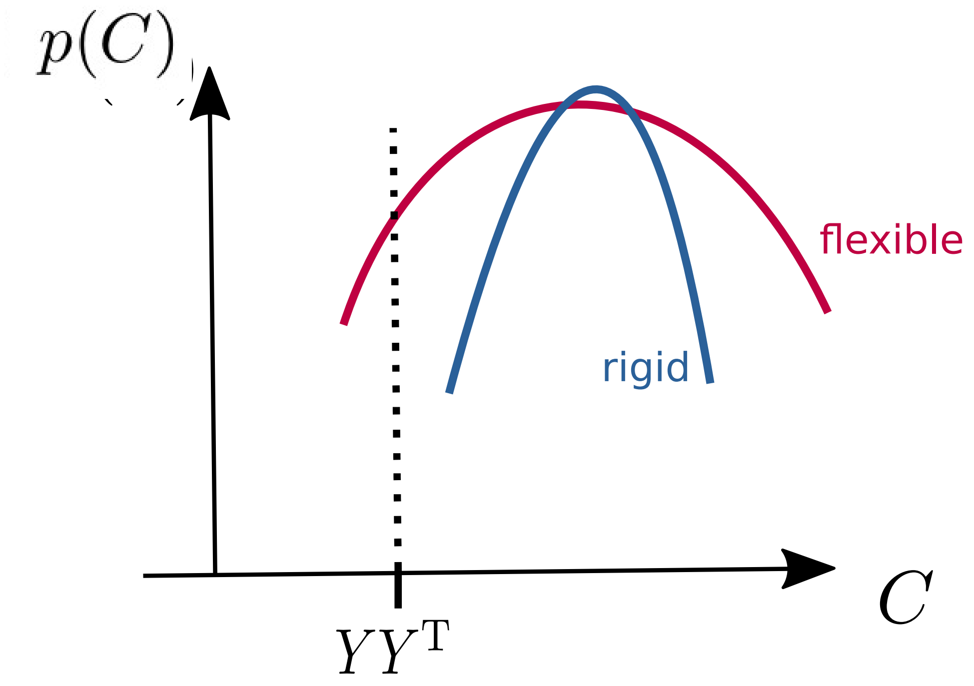




FLUCTUATIONS LEAD TO FEATURE LEARNING

prior

$$p(Y|X) = \int \mathcal{D}C\, \mathcal{N}\left(Y|0, C^{(L)} + \kappa\mathbb{I}\right)\, p(C)$$



WHAT ABOUT FINITE N?

theory

so far:

$$N \longrightarrow \infty \quad , \quad P = \alpha N \rightarrow \infty \quad \Big\} \text{ proportional limit}$$

tools:
- large deviation theory
- perturbation expansion $P \ll N$

now:

$$\boxed{\begin{aligned} N &= \text{finite} \\ P &= \text{finite} \end{aligned}}$$

tools:
- field theory
- fluctuation expansion around NNGP

numerics

$$\boxed{\begin{aligned} N &= \text{finite} \\ P &= \text{finite} \end{aligned}}$$

FLUCTUATION CORRECTIONS

network prior, keeping auxiliary field

$$(C, \tilde{C}) \sim \exp\left(\mathcal{S}(C, \tilde{C}) + \mathcal{S}_{\mathrm{D}}(C^{(L)}|Y)\right)$$

$$\mathcal{S}(C, \tilde{C}) := -\mathrm{tr}\,\tilde{C}^{\mathsf{T}} C + \mathcal{W}(\tilde{C}|C)$$

fluctuations around NNGP

$$C = C^{*}_{\mathrm{NNGP}} + \delta C$$

$$\tilde{C} = 0 + \delta\tilde{C}$$

fluctuation expansion (Gaussian fluctuations around NNGP)

$$(\delta C, \delta\tilde{C}) \sim \exp\left(\frac{1}{2}(\delta C, \delta\tilde{C})^{\mathsf{T}} \mathcal{S}^{(2)}\,(\delta C, \delta\tilde{C}) + \mathcal{S}_{\mathrm{D}}^{(1)\mathsf{T}} \delta C^{(L)}\right)$$

linear system of equations

$$\left[\mathcal{S}^{(2)} \begin{pmatrix} \delta C \\ \delta\tilde{C} \end{pmatrix}\right]^{(l)} + \begin{pmatrix} \mathcal{S}_{\mathrm{D}}^{(1)} \\ 0 \end{pmatrix} \delta_{l\,L} = 0$$

FLUCTUATION CORRECTIONS

theory

so far:

$$\left.\begin{array}{l} N \longrightarrow \infty \\ P = \alpha N \longrightarrow \infty \end{array}\right\}$$ proportional limit

tools:
- large deviation theory
- perturbation expansion $P \ll N$

now:

$$\boxed{\begin{array}{l} N = \text{finite} \\ P = \text{finite} \end{array}}$$

tools:
- field theory
- fluctuation expansion around NNGP

linearized forward mapping

$C^{(l-1)} \mapsto C^{(l)}$

$$\delta C_{\alpha\beta}^{(l)} = g_l \sum_{\gamma\delta} \frac{\partial \langle \phi_\alpha^{(l-1)} \phi_\beta^{(l-1)} \rangle_{\mathcal{N}(0,C^{(l-1)})}}{\partial C_{\gamma\delta}^{(l-1)}} \, \delta C_{\gamma\delta}^{(l-1)} + g_l^2 \sum_{\gamma\delta} V_{\alpha\beta,\gamma\delta}^{(l-1)} \, \delta \tilde{C}_{\gamma\delta}^{(l)}$$

same form of backward mapping

$\tilde{C}^{(l+1)} \mapsto \tilde{C}^{(l)}$

$$\tilde{C}_{\alpha\beta}^{(l)} = \mathcal{G}_{\alpha\beta} \, \tilde{C}_{\alpha\beta}^{(l+1)} + \delta_{\alpha\beta} \sum_{\gamma} \mathcal{H}_{\gamma\alpha} \tilde{C}_{\gamma\alpha}^{(l+1)} + \mathcal{O}(\tilde{C}^2)$$

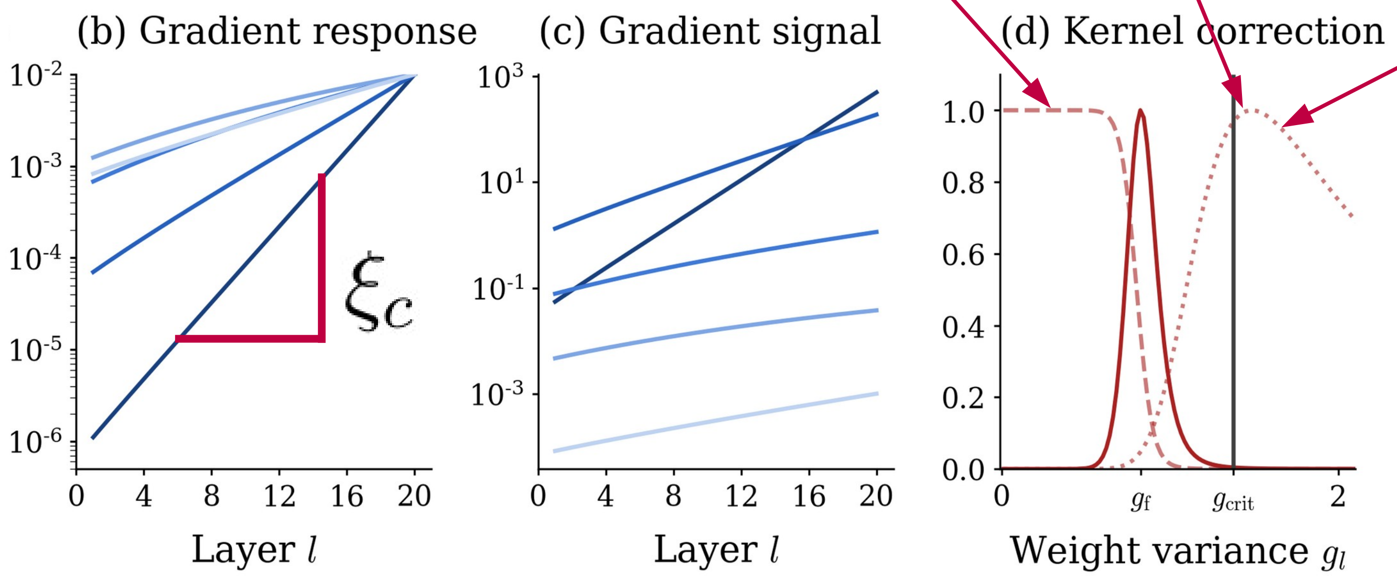
$$\mathcal{G}_{\alpha\beta} \, , \mathcal{H}_{\alpha\beta}, V_{\alpha\beta,\gamma\delta} \big|_{C^*_{\mathrm{NNGP}}}$$

FEATURE LEARNING CLOSE TO CRITICALITY

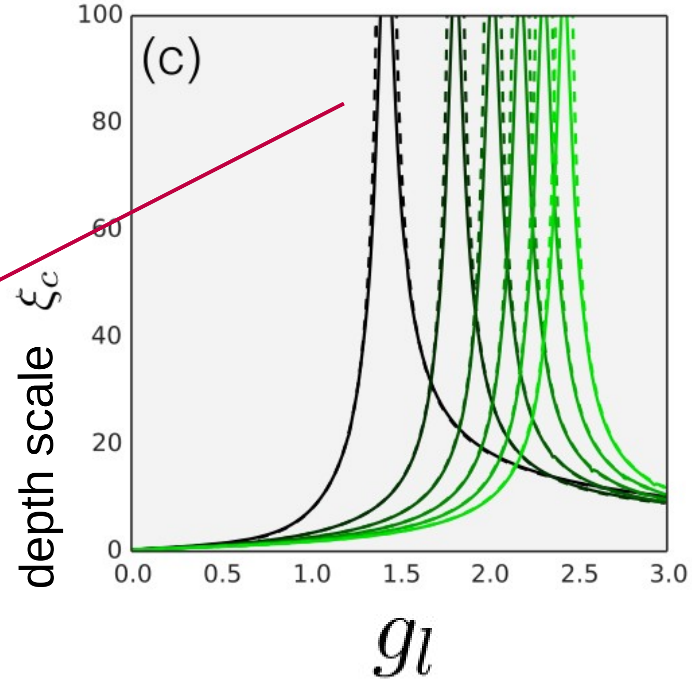
Interplay between backward response function and error signal in output layer

$$\tilde{C}^{(l)}_{\alpha\beta} = \tilde{C}^{(L)}_{\alpha\beta} \prod_{s=l}^{L-1} g_{s+1} \left\langle \left(\phi^{(s)}_\alpha\right)' \left(\phi^{(s)}_\beta\right)' \right\rangle_{\mathcal{N}(0,C^{(s)})}$$

$$\tilde{C}^{(L)} = \frac{\partial}{\partial C^{(L)}} \ln \mathcal{N}(y|0, C^{(L)} + \kappa \mathbb{I})$$



Schoenholz et al. 2017



divergence at transition to chaos