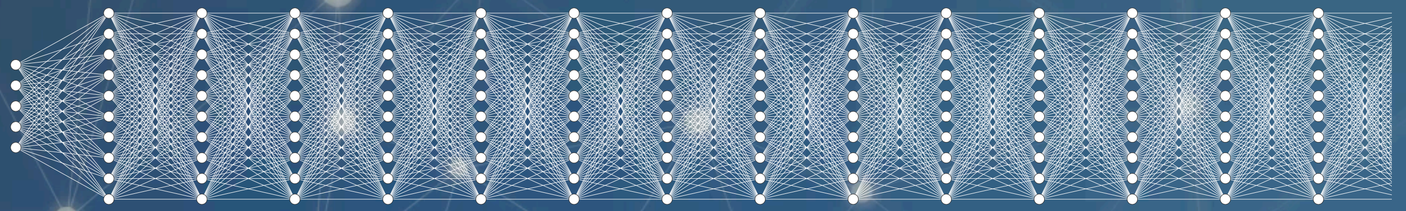
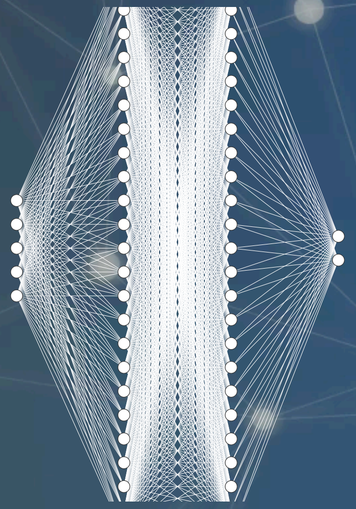


# On the Infinite Width and Depth Limits of Predictive Coding Networks



Francesco Innocenti,
El Mehdi Achour & Rafal Bogacz

UNIVERSITY OF OXFORD

ICML
International Conference On Machine Learning

UM6P

College of Computing

# Overview

1. Introduction
2. Main results
3. Discussion

# Overview

1. **Introduction**
2. Main results
3. Discussion

# Introduction

blue brain

- Backpropagation (BP) is the core algorithm for training neural networks, yet it is energy inefficient and unlikely to be implemented by the brain because of its non-local nature
- Predictive coding (PC) is an influential "biologically plausible" learning algorithm based on the basic idea that neurons minimise their prediction errors
- Despite some recent encouraging progress, training wide and especially deep PC networks (PCNs) on large-scale datasets competitively with BP remains an open challenge
- Inspired by the recent sucess of analysing the infinite width and depth limits of BP-trained networks, we theoretically derive and empirically validate the first stable parameterisations for wide and deep PCNs, including convolutional nets and transformers

# TL;DR

The stable and “rich” (non-lazy) parameterisations in model width and depth for predictive coding (PC) turn out to be the exactly same as for backprop (BP).

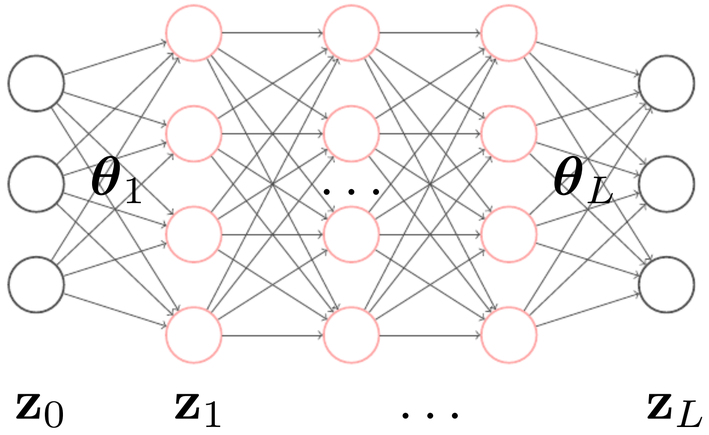
Under any of these parameterisations, the weight gradients computed by PC converge to BP’s in much wider than deep networks like the brain.

# Predictive coding networks (PCNs)

- PCNs minimise an energy function that is a sum of layer-wise prediction errors

$$\mathcal{F}(\mathbf{z}, \boldsymbol{\theta}) = \sum_{\ell=1}^{L} || \underbrace{\mathbf{z}_\ell - f_\ell(\mathbf{z}_{\ell-1}; \boldsymbol{\theta}_\ell)}_{\text{layer prediction error}} ||_2^2$$

$$\mathcal{L}(\boldsymbol{\theta}) = ||\mathbf{y} - f\ (\mathbf{x}; \boldsymbol{\theta})||_2^2$$

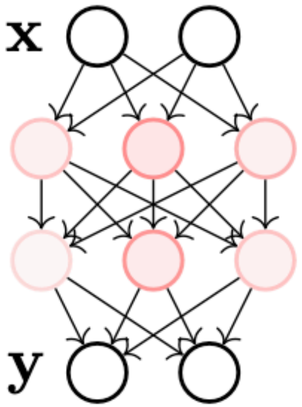


# Predictive coding networks

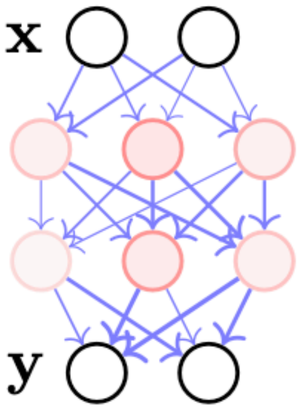
$$\mathcal{F}(\mathbf{z}, \boldsymbol{\theta}) = \sum_{\ell=1}^{L} || \underbrace{\mathbf{z}_\ell - f_\ell(\mathbf{z}_{\ell-1}; \boldsymbol{\theta}_\ell)}_{\text{layer prediction error}} ||_2^2$$

- In supervised learning, we clamp the first and last layers to some input and target data

- PCNs are then trained by minimising the energy in two alternating phases:




*Infer:* $\mathbf{z}^* = \arg\min_{\mathbf{z}} \mathcal{F}(\boldsymbol{\theta}_t, \mathbf{z})$




*Learn:* $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \mathbf{P}_t \nabla_\theta \mathcal{F}(\boldsymbol{\theta}_t, \mathbf{z}^*)$

- Note that both the activity and weight gradients of PC are local, in the sense that they require only information from neighbouring layers

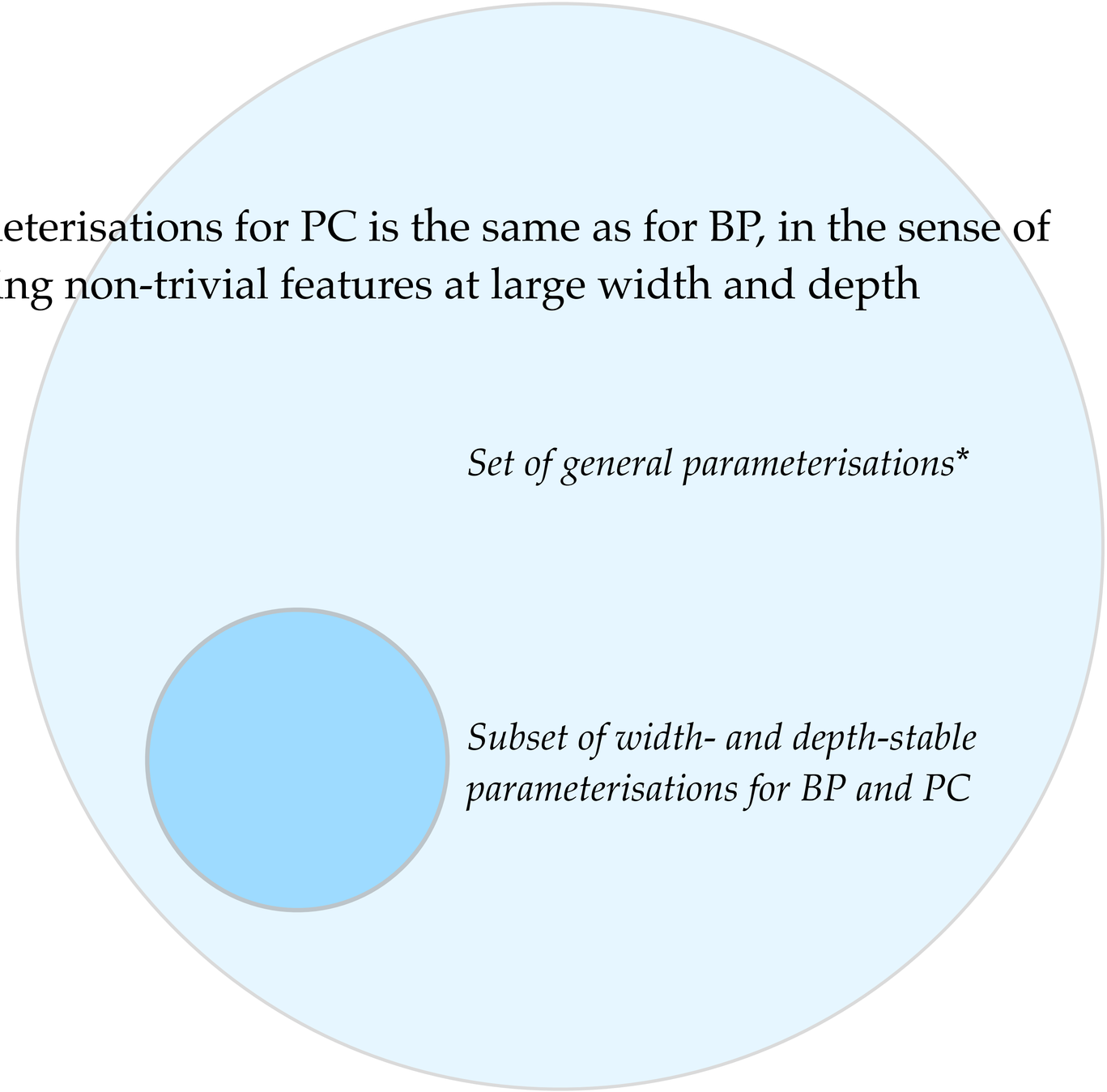
# Overview

1. Introduction
2. **Main results**
3. Discussion

# Main results

- (*Informal*) The set of scalable parameterisations for PC is the same as for BP, in the sense of being numerically stable and learning non-trivial features at large width and depth

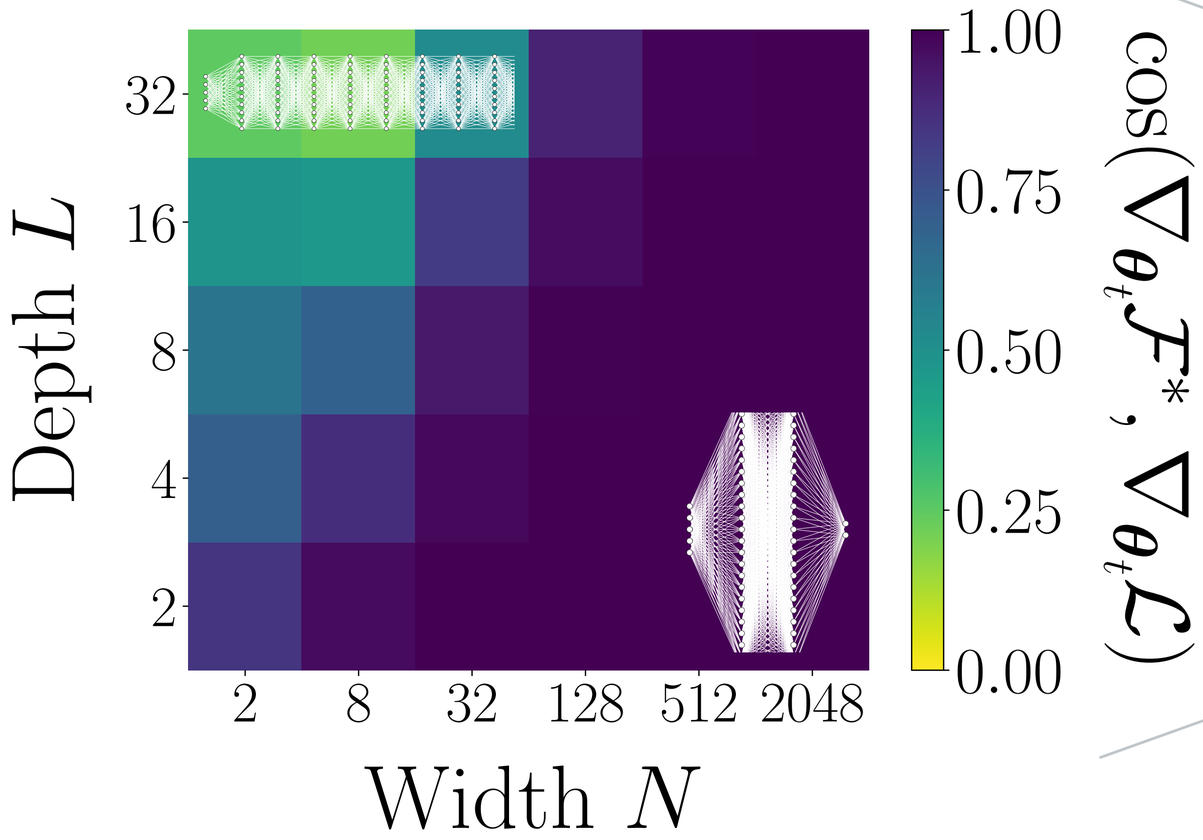
**Theorem 2.** *(Width- and depth-stable feature-learning parameterisations for linear PCNs.) Consider any width-stable and feature-learning parameterisation of linear ResNets (Eq. 15) with an additional depth scaling exponent $\alpha$ (Eqs. 7-8, 18 & 10). Assume PCNs that learn on the equilibrated energy (with rescaling as in Eq. 104). Then, the parameterisation that satisfies the depth Desiderata for BP is the same as for PC (i.e. $\alpha = 1/2$).*

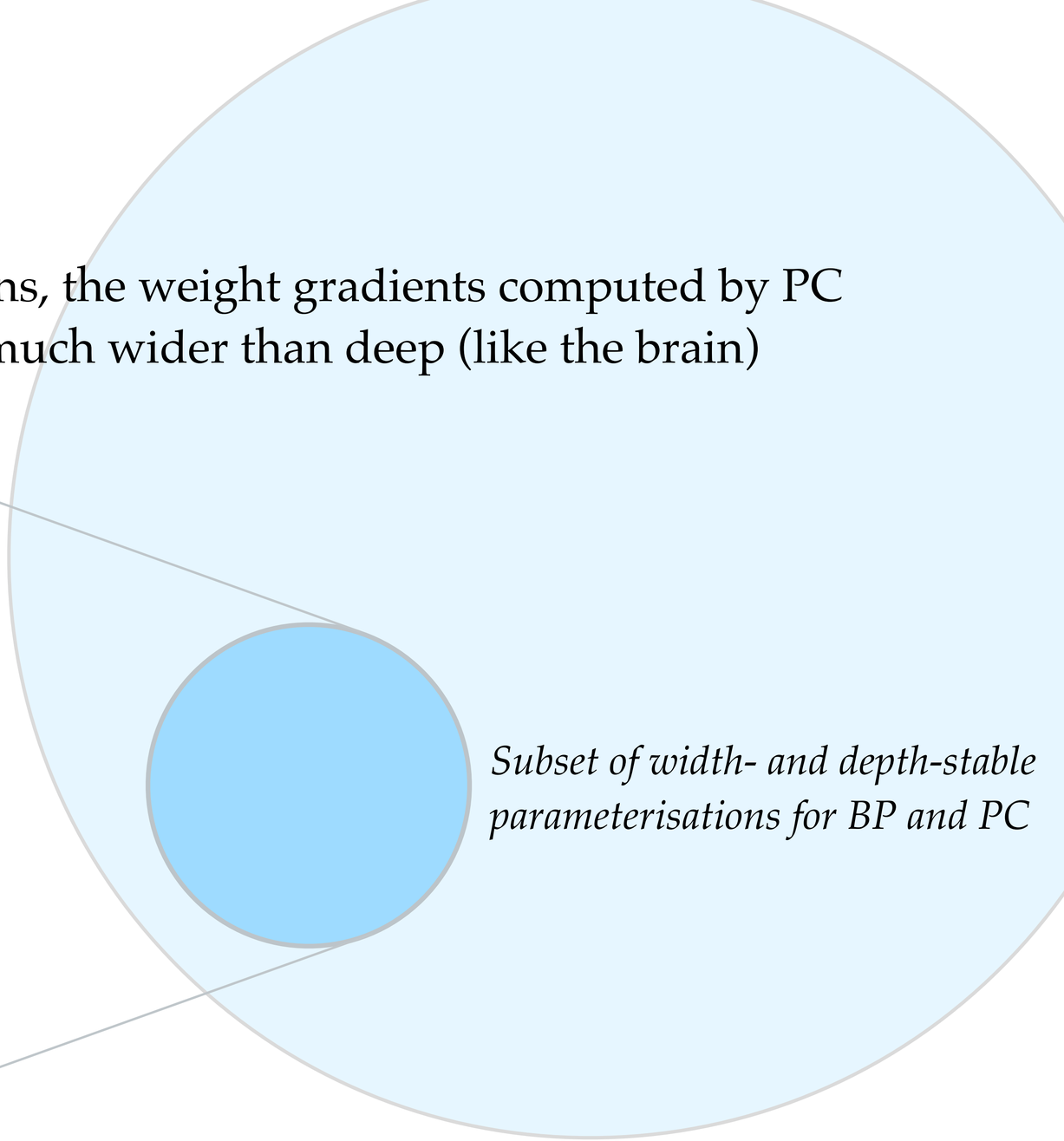



*Not all possible parameterisations

# Main results

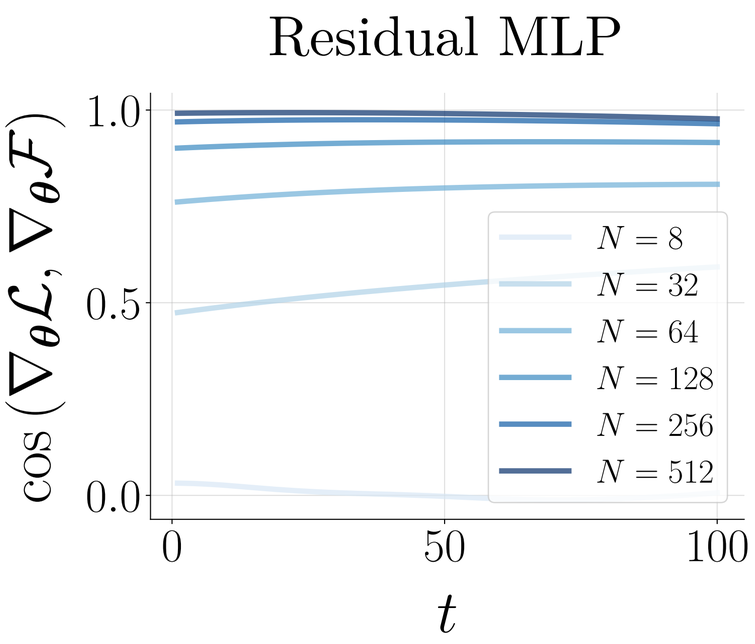
- (*Corollary*) Under any of these parameterisations, the weight gradients computed by PC converge to backprop's for networks that are much wider than deep (like the brain)

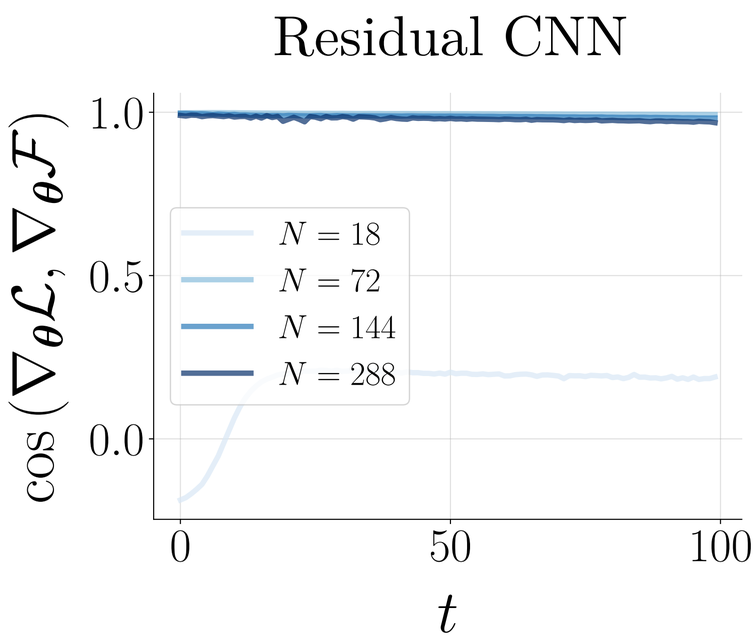


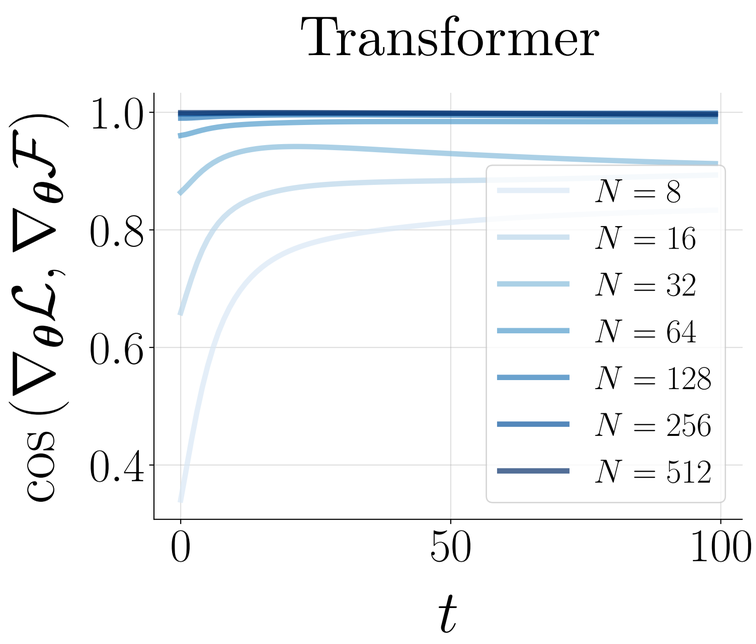



# Main results

- **This result is very general**: it empirically holds for nonlinear networks including MLPs, ResNets, CNNs, (nano-GPT-style) transformers, trained with different optimisers (e.g. Adam) and loss functions (e.g. CE), on small and large-scale datasets (e.g. ImageNet)



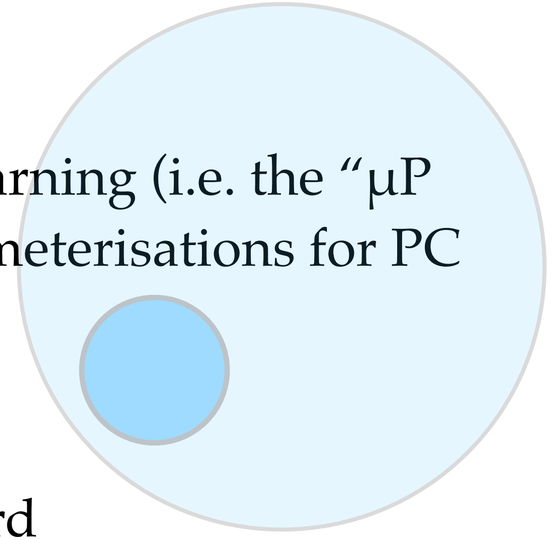



# Overview

1. Introduction
2. Main results
3. **Discussion**

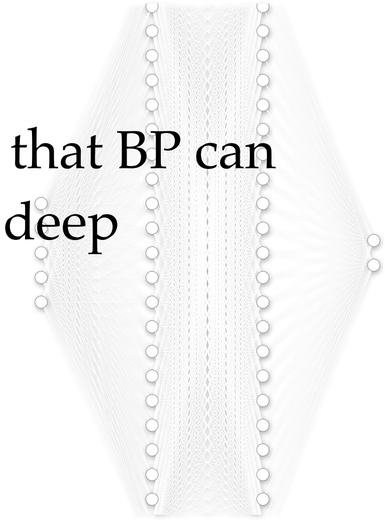
# Main implications



- **If** one would like to satisfy reasonable notions of stability and feature-learning (i.e. the "μP desiderata"), **then necessarily** the set of width- and depth-scalable parameterisations for PC is the same as for BP

- Importantly, this means that **PCNs trained in practice** (with the "standard parameterisation") **cannot be stably scaled in width and depth**, including regimes where PC has shown benefits over BP

- At the same time, our results are the first (to the best of our knowledge) to show that BP can be effectively implemented with a local algorithm **at scale**, for much wider than deep networks like the brain



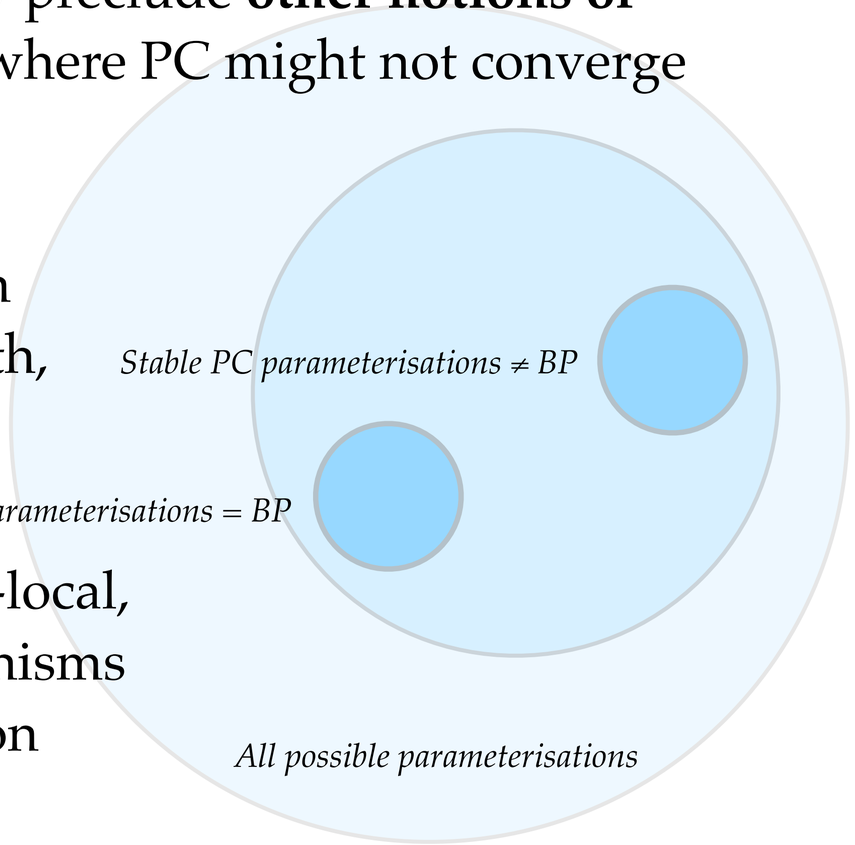
# Limitations & future directions

- **Linear theory**: It could be interesting to see if one could generalise our theory to nonlinear models using tools from statistical physics (i.e. dynamical mean field theory)

- **Alternative parameterisations**: Our work does not necessarily preclude **other notions or desiderata** of a stable and feature-learning parameterisation, where PC might not converge to (and perhaps be better than) BP in some limit

- **PC inference cost**: If we could accelerate this (most likely with analog hardware), training could be sped up by a factor ~depth, since weight updates are parallelisable across layers with PC

- **Bio-plausible attention**: Standard self-attention is highly non-local, and it could be interesting to study more bio-plausible mechanisms where the softmax is itself the gradient of some energy function



# Code

https://github.com/thebuckleylab/jpc/tree/main/experiments/limits_paper

README.md

## On the ∞ Width & Depth Limits of PCNs

Paper arXiv:2508.01191

Code for the paper On the Infinite Width and Depth Limits of Predictive Coding Networks.

### Setup

Clone the `jpc` repo. We recommend using a virtual environment, e.g.

```
python3 -m venv venv
```

Install `jpc`

```
pip install -e .
```

# *Thank you for your attention!*

El Mehdi Achour

Rafal Bogacz

UNIVERSITY OF OXFORD

ICML International Conference On Machine Learning

UM6P College of Computing