MACHINE LEARNING FOR LATTICE FIELD THEORY AND BEYOND — TRENTO 30/06/2023

STATISTICAL MECHANICS OF DEEP LEARNING BEYOND THE INFINITE-WIDTH LIMIT

Pietro Rotondo — University of Parma

UNIVERSITÀ
DI PARMA

[arXiv:2209.04882 (2022)]

OUTLINE OF THE TALK

OUTLINE OF THE TALK

- Overparametrised DNNs / generalisation / Stat. phys. approaches

OUTLINE OF THE TALK

- Overparametrised DNNs / generalisation / Stat. phys. approaches

- Two important ideas for this work:
  - ① Infinite-width limit
  - ② Statistical mechanics of deep linear networks

OUTLINE OF THE TALK

- Overparametrised DNNs / generalisation / Stat. phys. approaches

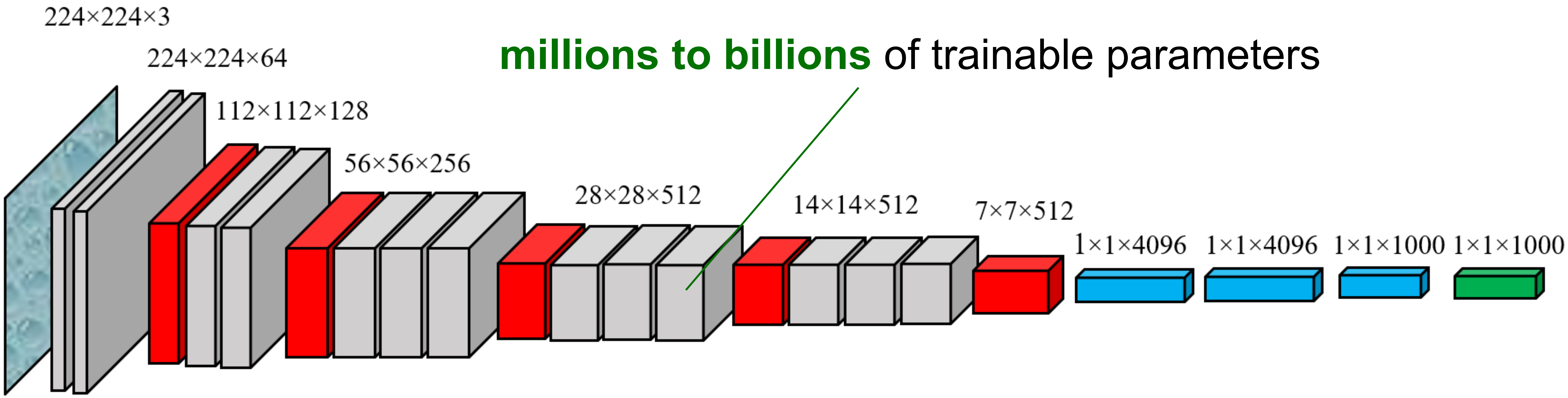
- Two important ideas for this work:

    ① Infinite-width limit

    ② Statistical mechanics of deep linear networks

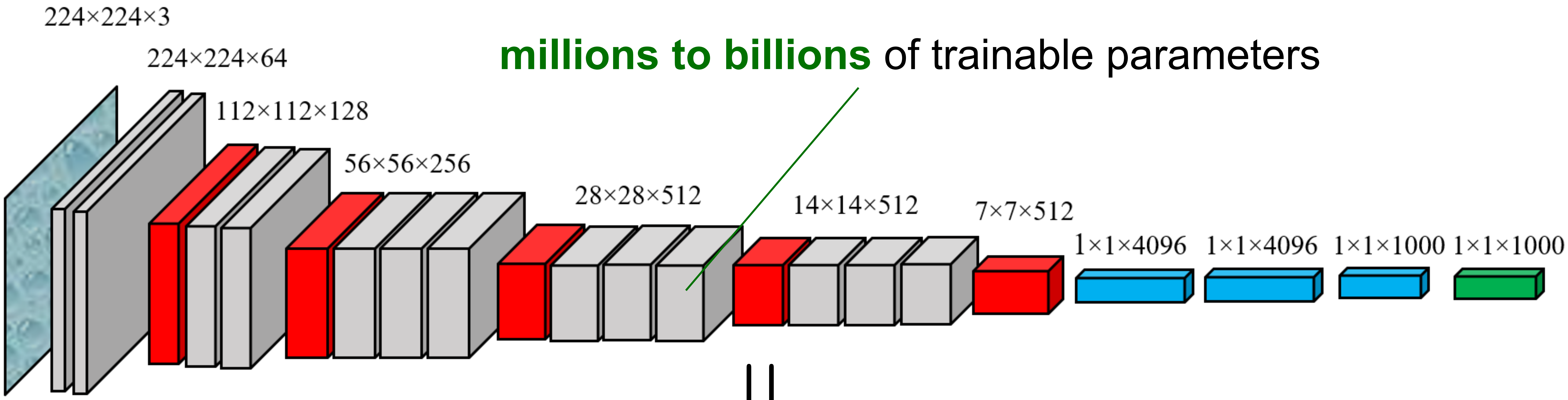
- Results: an analytical framework to investigate the partition function of DNNs at “finite width”

OVERPARAMETRISATION IN DEEP NETS: A BLESS FOR PRACTITIONERS, A PROBLEM FOR THEORISTS



millions to billions of trainable parameters

OVERPARAMETRISATION IN DEEP NETS: A BLESS FOR PRACTITIONERS, A PROBLEM FOR THEORISTS



millions to billions of trainable parameters

$$f_{\mathrm{DNN}}(\mathbf{x}) = v \circ \sigma \circ W^{(L)} \circ \sigma \circ W^{(L-1)} \circ \cdots \circ \sigma \circ W^{(1)}(\mathbf{x})$$

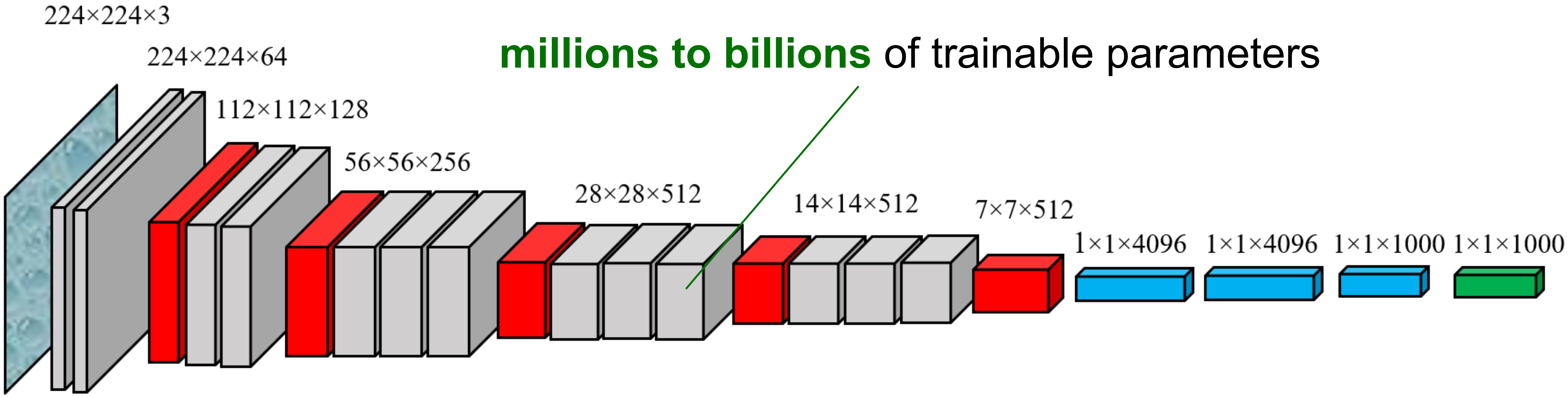
non-linear activation function

affine transformation

$$f : \mathbb{R}^N \to \mathbb{R}$$

N is typically large

OVERPARAMETRISATION IN DEEP NETS: A BLESS FOR PRACTITIONERS, A PROBLEM FOR THEORISTS



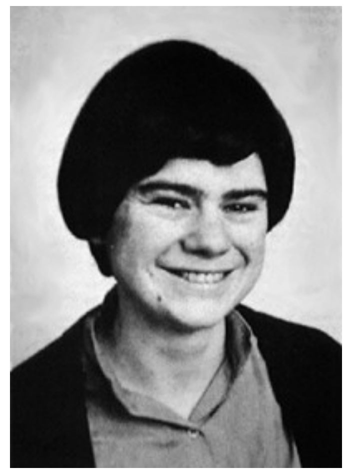
millions to billions of trainable parameters

overparametrised regime

number of trainable parameters $\gg$ **size of the training set**

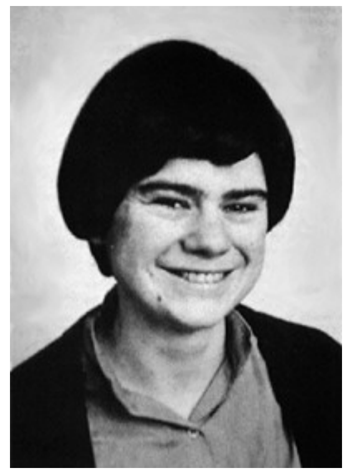
$L \times N^2$ P

STATISTICAL MECHANICS APPROACHES ARE LIMITED TO VERY SIMPLE ARCHITECTURES (FOR THE MOMENT)



E. Gardner

H. Sompolinsky



E. Gardner

H. Sompolinsky

STATISTICAL MECHANICS APPROACHES ARE LIMITED TO VERY SIMPLE ARCHITECTURES (FOR THE MOMENT)

what we would like to investigate

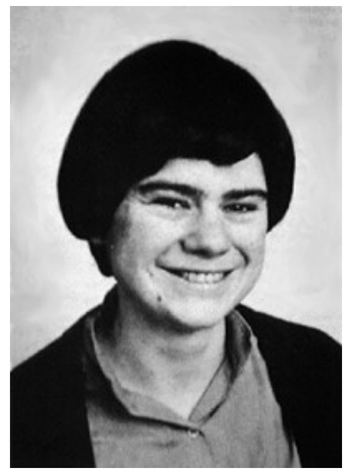
Deep neural networks
(fully-connected and/or convolutional)

vs

what we know how to investigate

Linear model (perceptron)
Random feature model
Kernel learning (SVMs)
Committee machine

STATISTICAL MECHANICS APPROACHES ARE LIMITED TO VERY SIMPLE ARCHITECTURES (FOR THE MOMENT)



E. Gardner

H. Sompolinsky

what we would like to investigate

Deep neural networks
(fully-connected and/or convolutional)

vs

what we know how to investigate

Linear model (perceptron)
Random feature model
Kernel learning (SVMs)
Committee machine

$$\mathcal{T} = \{(\mathbf{x}^\mu, y^\mu)\}_{\mu=1}^P \text{———} P(\mathbf{x}, y)$$

input-output probability distribution

training set

$$Z = \int \mathcal{D}\theta \, e^{-\beta \sum_{\mu=1}^P \ell(y^\mu, f_\theta(\mathbf{x}^\mu))}$$

partition function

$$\langle \log Z \rangle = \lim_{n \to 0} \frac{\langle Z^n \rangle - 1}{n}$$

replica trick

WHAT DO WE KNOW? (1) THE INFINITE-WIDTH LIMIT OF DEEP NEURAL NETWORKS

[R. M. Neal, “Bayesian Learning for Neural Networks”, Springer (1996)]
[J. Lee et al., ICLR (2018)] [A. Jacot, F. Gabriel, C. Hongler, NeurIPS (2018)]

$$\boxed{N_\ell \gg P \quad \forall\, \ell = 1, \ldots, L}$$

WHAT DO WE KNOW? (1) THE INFINITE-WIDTH LIMIT OF DEEP NEURAL NETWORKS

[R. M. Neal, “Bayesian Learning for Neural Networks”, Springer (1996)]
[J. Lee et al., ICLR (2018)] [A. Jacot, F. Gabriel, C. Hongler, NeurIPS (2018)]

$$\boxed{N_\ell \gg P \quad \forall\, \ell = 1, \dots, L}$$

- Infinite-width deep neural networks are equivalent to Gaussian processes

$$K_\ell(\mathbf{x}_1, \mathbf{x}_2) = \int dz_1 dz_2\, \mathcal{N}\left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{\ell-1}(\mathbf{x}_1, \mathbf{x}_1) & K_{\ell-1}(\mathbf{x}_1, \mathbf{x}_2) \\ K_{\ell-1}(\mathbf{x}_2, \mathbf{x}_1) & K_{\ell-1}(\mathbf{x}_2, \mathbf{x}_2) \end{bmatrix} \right) \sigma(z_1)\sigma(z_2) \qquad K_0(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{N_0}$$

WHAT DO WE KNOW? (1) THE INFINITE-WIDTH LIMIT OF DEEP NEURAL NETWORKS

[R. M. Neal, “Bayesian Learning for Neural Networks”, Springer (1996)]
[J. Lee et al., ICLR (2018)] [A. Jacot, F. Gabriel, C. Hongler, NeurIPS (2018)]

$$\boxed{N_\ell \gg P \quad \forall\, \ell = 1, \dots, L}$$

Bayesian vs Gradient Descent:
NNGP vs NTK!!

- Infinite-width deep neural networks are equivalent to Gaussian processes

$$K_\ell(\mathbf{x}_1, \mathbf{x}_2) = \int dz_1 dz_2 \, \mathcal{N}\left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{\ell-1}(\mathbf{x}_1, \mathbf{x}_1) & K_{\ell-1}(\mathbf{x}_1, \mathbf{x}_2) \\ K_{\ell-1}(\mathbf{x}_2, \mathbf{x}_1) & K_{\ell-1}(\mathbf{x}_2, \mathbf{x}_2) \end{bmatrix} \right) \sigma(z_1)\sigma(z_2) \qquad K_0(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{N_0}$$

WHAT DO WE KNOW? (1) THE INFINITE-WIDTH LIMIT OF DEEP NEURAL NETWORKS

[R. M. Neal, “Bayesian Learning for Neural Networks”, Springer (1996)]
[J. Lee et al., ICLR (2018)] [A. Jacot, F. Gabriel, C. Hongler, NeurIPS (2018)]

$$\boxed{N_\ell \gg P \quad \forall\, \ell = 1, \dots, L}$$

Bayesian vs Gradient Descent:
NNGP vs NTK!!

- Infinite-width deep neural networks are equivalent to Gaussian processes

$$K_\ell(\mathbf{x}_1, \mathbf{x}_2) = \int dz_1 dz_2\, \mathcal{N}\left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{\ell-1}(\mathbf{x}_1, \mathbf{x}_1) & K_{\ell-1}(\mathbf{x}_1, \mathbf{x}_2) \\ K_{\ell-1}(\mathbf{x}_2, \mathbf{x}_1) & K_{\ell-1}(\mathbf{x}_2, \mathbf{x}_2) \end{bmatrix} \right) \sigma(z_1)\sigma(z_2) \qquad K_0(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{N_0}$$

- Data-averaged partition functions can be studied in this limit

[A. Canatar, B. Bordelon, C. Pehlevan, Nat. Comm. (2020)]
[R. Dietrich, M. Opper, H. Sompolinsky, PRL (1999)]

WHAT DO WE KNOW? (2) STATISTICAL MECHANICS OF DEEP LINEAR NETWORKS

[Q. Li & H. Sompolinsky, PRX (2021)] [A. Saxe, J. McClelland, S. Ganguli, ICLR (2014)]

$$f_{\mathrm{DLN}}(\mathbf{x}) = v \circ \xcancel{\sigma} \circ W^{(L)} \circ \xcancel{\sigma} \circ W^{(L-1)} \circ \cdots \circ \xcancel{\sigma} \circ W^{(1)}(\mathbf{x})$$

$$Z = \int \mathcal{D}\theta \, e^{-\beta \sum_{\mu=1}^{P} \ell(y^\mu, f_\theta(\mathbf{x}^\mu))}$$

$$\ell(y^\mu, f_\theta(\mathbf{x}^\mu)) = (y^\mu - f_\theta(\mathbf{x}^\mu))^2$$

WHAT DO WE KNOW? (2) STATISTICAL MECHANICS OF DEEP LINEAR NETWORKS

[Q. Li & H. Sompolinsky, PRX (2021)] [A. Saxe, J. McClelland, S. Ganguli, ICLR (2014)]

$$f_{\mathrm{DLN}}(\mathbf{x}) = v \circ \xcancel{\sigma} \circ W^{(L)} \circ \xcancel{\sigma} \circ W^{(L-1)} \circ \cdots \circ \xcancel{\sigma} \circ W^{(1)}(\mathbf{x})$$

$$Z = \int \mathcal{D}\theta \, e^{-\beta \sum_{\mu=1}^{P} \ell(y^\mu, f_\theta(\mathbf{x}^\mu))} \qquad \ell(y^\mu, f_\theta(\mathbf{x}^\mu)) = (y^\mu - f_\theta(\mathbf{x}^\mu))^2$$

- **IDEA: integrate the weights backwards, starting from the output layer!**

(Backpropagating kernel renormalisation)

$$u_\ell \qquad \ell = 1, \ldots, L \qquad\qquad P\,, N_\ell \to \infty \qquad \alpha_\ell = \frac{P}{N_\ell}$$

determined self-consistently

Thermodynamic limit

WHAT DO WE KNOW? (2) STATISTICAL MECHANICS OF DEEP LINEAR NETWORKS

[Q. Li & H. Sompolinsky, PRX (2021)] [A. Saxe, J. McClelland, S. Ganguli, ICLR (2014)]

$$f_{\mathrm{DLN}}(\mathbf{x}) = v \circ \cancel{\sigma} \circ W^{(L)} \circ \cancel{\sigma} \circ W^{(L-1)} \circ \cdots \circ \cancel{\sigma} \circ W^{(1)}(\mathbf{x})$$

average generalisation error over a new unseen example

isotropic limit

$$\alpha_\ell = \alpha = \frac{P}{N}$$

linear kernel

$$(K_0)_{\mu\nu} = \frac{\mathbf{x}^\mu \cdot \mathbf{x}^\nu}{N_0}$$

$$\langle \epsilon_{\mathrm{g}}(\mathbf{x}^0, y^0) \rangle = \left[y^0 - \sum_{\mu\nu} \kappa_\mu(\mathbf{x}^0)(K_0^{-1})_{\mu\nu}\, y_\nu, \right]^2 + u_0^L \left[\kappa_0(\mathbf{x}^0) - \sum_{\mu\nu} \kappa_\mu(\mathbf{x}^0)(K_0^{-1})_{\mu\nu}\, \kappa_\nu(\mathbf{x}^0) \right]$$

$$r_0 = \frac{\sigma^{2L}}{P} y^T K_0^{-1} y$$

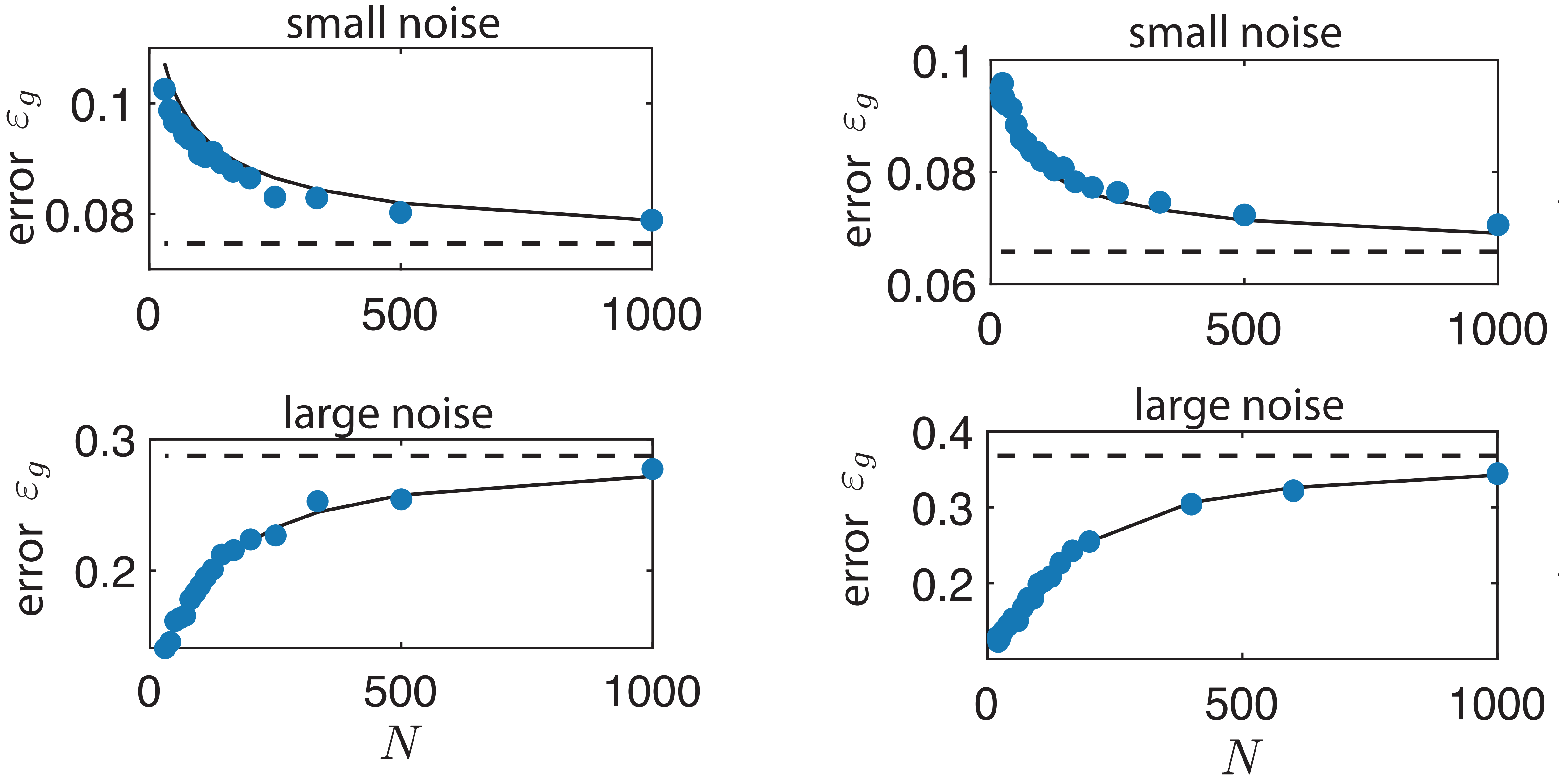
$$1 - \frac{u_0}{\sigma^2} = \alpha \left(1 - \frac{r_0}{u_0^L} \right)$$

WHAT DO WE KNOW? (3) AN HEURISTIC THEORY FOR RELU ACTIVATION

[Q. Li & H. Sompolinsky, PRX (2021)]

- **IDEA! Replace the linear kernel with the nonlinear kernel for ReLU activation**

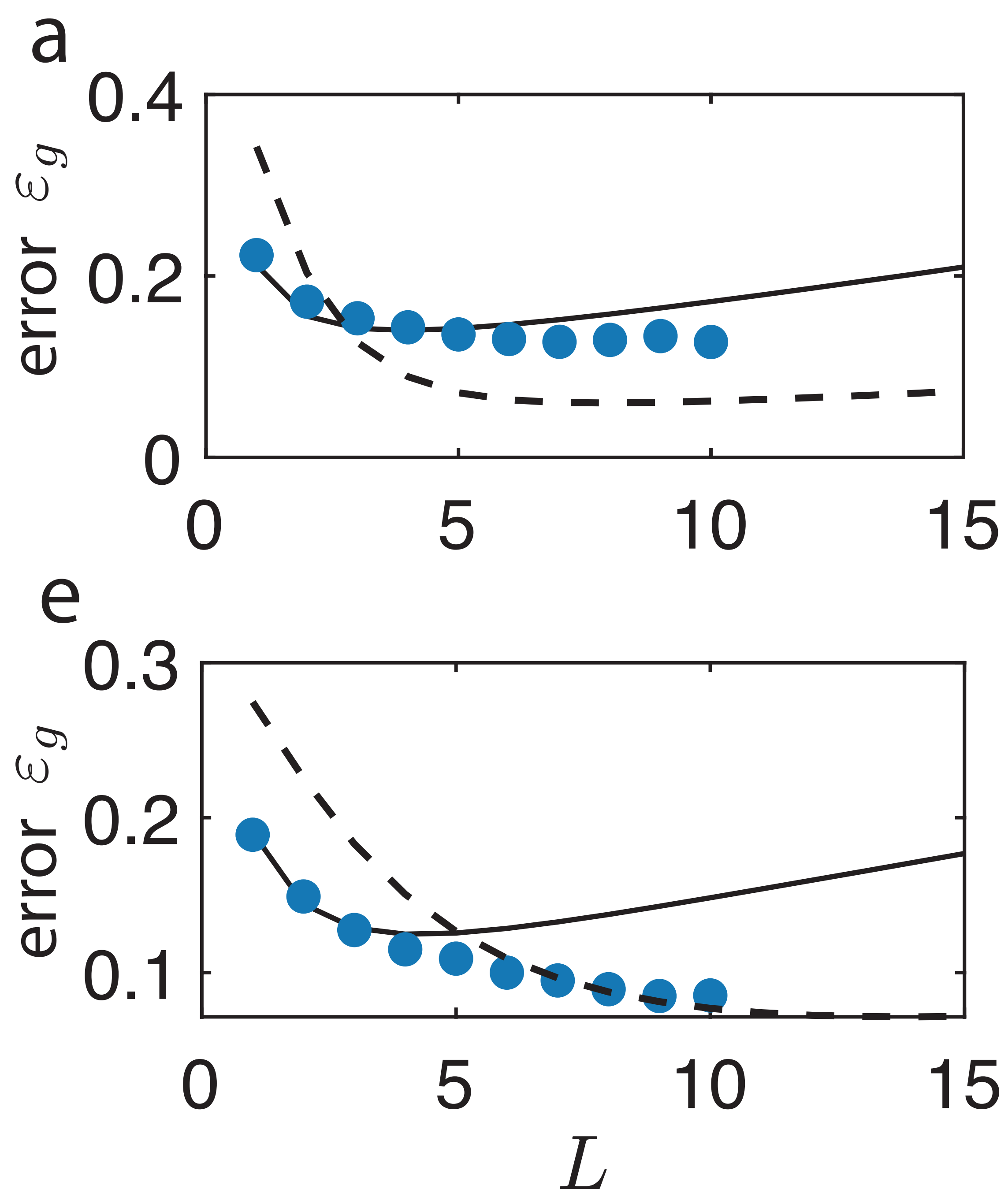
1-hidden layer (ReLU)



binary classification MNIST

noisy linear teacher

L-hidden layer (ReLU)



MAIN GOAL: developing an analytical framework based on statistical mechanics to describe deep learning beyond the infinite-width limit

① $P\,, N_\ell \to \infty \qquad \alpha_\ell = \dfrac{P}{N_\ell} \qquad \ell = 1, \dots, L$ (Thermodynamic limit)

② Fixed instance of the training set $\mathcal{T} = \{(\mathbf{x}^\mu, y^\mu)\}_{\mu=1}^{P}$

SETTING OF THE LEARNING PROBLEM

- $$h^{(\ell)}_{i_\ell} = \frac{1}{\sqrt{N_{\ell-1}}} \sum_{i_{\ell-1}=1}^{N_{\ell-1}} W^{(\ell)}_{i_\ell i_{\ell-1}} \, \sigma\left(h^{(\ell-1)}_{i_{\ell-1}}\right) + b^{(\ell)}_{i_\ell} \,,$$

  pre-activations at each layer

  $$h^{(1)}_{i_1} = \frac{1}{\sqrt{N_0}} \sum_{i_0=1}^{N_0} W^{(1)}_{i_1 i_0} x_{i_0} + b^{(1)}_{i_1}$$

  $$f_{\mathrm{DNN}}(\mathbf{x}) = \frac{1}{\sqrt{N_L}} \sum_{i_L=1}^{N_L} v_{i_L} \sigma\left[h^{(L)}_{i_L}(\mathbf{x})\right]$$

  readout layer

- $$\mathcal{L} = \frac{1}{2} \sum_{\mu=1}^{P} \left[y^\mu - f_{\mathrm{DNN}}(\mathbf{x}^\mu)\right]^2 + \mathcal{L}_{\mathrm{reg}} \,,$$

  $$\mathcal{L}_{\mathrm{reg}} = \frac{\lambda_L}{2\beta} \sum_{i_L=1}^{N_L} v_{i_L}^2 + \frac{1}{2\beta} \sum_{\ell=0}^{L-1} \lambda^{(\ell)} \|W^{(\ell)}\|^2$$

regression problem
quadratic loss function

$$\mathcal{T} = \{(\mathbf{x}^\mu, y^\mu)\}_{\mu=1}^{P}$$

A BAYESIAN DESCRIPTION OF LEARNING (AKA EQUILIBRIUM STATISTICAL MECHANICS)

$$Z = \int \mathcal{D}\theta \, e^{-\beta \mathcal{L}(\theta)}$$

$$\mathcal{L} = \frac{1}{2} \sum_{\mu=1}^{P} \left[y^{\mu} - f_{\mathrm{DNN}}(\mathbf{x}^{\mu}) \right]^2 + \mathcal{L}_{\mathrm{reg}} \, ,$$

$$\mathcal{L}_{\mathrm{reg}} = \frac{\lambda_L}{2\beta} \sum_{i_L=1}^{N_L} v_{i_L}^2 + \frac{1}{2\beta} \sum_{\ell=0}^{L-1} \lambda^{(\ell)} \|W^{(\ell)}\|^2$$

quadratic loss function

A BAYESIAN DESCRIPTION OF LEARNING (AKA EQUILIBRIUM STATISTICAL MECHANICS)

$$\boxed{Z} = \int \mathcal{D}\theta \, e^{-\beta \mathcal{L}(\theta)}$$

linked to the posterior distribution of the weights after training

$$\mathcal{L} = \frac{1}{2}\sum_{\mu=1}^{P}\left[y^\mu - f_{\mathrm{DNN}}(\mathbf{x}^\mu)\right]^2 + \mathcal{L}_{\mathrm{reg}}\,,$$

$$\mathcal{L}_{\mathrm{reg}} = \frac{\lambda_L}{2\beta}\sum_{i_L=1}^{N_L} v_{i_L}^2 + \frac{1}{2\beta}\sum_{\ell=0}^{L-1}\lambda^{(\ell)}\|W^{(\ell)}\|^2$$

quadratic loss function

A BAYESIAN DESCRIPTION OF LEARNING (AKA EQUILIBRIUM STATISTICAL MECHANICS)

$$Z = \int \mathcal{D}\theta \, e^{-\beta \mathcal{L}(\theta)}$$

linked to the posterior distribution of the weights after training

$$\mathcal{L} = \frac{1}{2} \sum_{\mu=1}^{P} \left[y^{\mu} - f_{\mathrm{DNN}}(\mathbf{x}^{\mu}) \right]^2 + \mathcal{L}_{\mathrm{reg}} \, ,$$

$$\mathcal{L}_{\mathrm{reg}} = \frac{\lambda_L}{2\beta} \sum_{i_L=1}^{N_L} v_{i_L}^2 + \frac{1}{2\beta} \sum_{\ell=0}^{L-1} \lambda^{(\ell)} \| W^{(\ell)} \|^2$$

quadratic loss function

Regularisation should be interpreted as a Gaussian prior over the weights

A BAYESIAN DESCRIPTION OF LEARNING (AKA EQUILIBRIUM STATISTICAL MECHANICS)

$$\boxed{Z} = \int \mathcal{D}\theta \, e^{-\beta \mathcal{L}(\theta)}$$

linked to the posterior distribution of the weights after training

$$\mathcal{L} = \frac{1}{2} \sum_{\mu=1}^{P} \left[y^\mu - f_{\mathrm{DNN}}(\mathbf{x}^\mu) \right]^2 + \mathcal{L}_{\mathrm{reg}} \, ,$$

$$\mathcal{L}_{\mathrm{reg}} = \frac{\lambda_L}{2\beta} \sum_{i_L=1}^{N_L} v_{i_L}^2 + \frac{1}{2\beta} \sum_{\ell=0}^{L-1} \lambda^{(\ell)} \| W^{(\ell)} \|^2$$

quadratic loss function

Regularisation should be interpreted as a Gaussian prior over the weights

$$\langle O(\theta) \rangle = \frac{1}{Z} \int \mathcal{D}\theta \; O(\theta) e^{-\beta \mathcal{L}(\theta)}$$

average of a generic observable of the weights

$$\epsilon_{\mathrm{g}}(\mathbf{x}^0, y^0; \theta) = \left(y^0 - f_\theta(\mathbf{x}^0) \right)^2$$

generalisation error

PARTITION FUNCTION FOR ONE HIDDEN LAYER NN IN THE ASYMPTOTIC LIMIT (1)

$$Z = \int \prod_{i_1}^{N_1} dv_{i_1} \prod_{i_1, i_0}^{N_1, N_0} dw_{i_1 i_0} \, \exp\left\{ -\frac{\lambda_1}{2} \sum_{i_1}^{N_1} v_{i_1}^2 - \frac{\lambda_0}{2} \|w\|^2 - \frac{\beta}{2} \sum_{\mu}^{P} \left[y^\mu - \frac{1}{\sqrt{N_1}} \sum_{i_1}^{N_1} v_{i_1} \sigma \left(\sum_{i_0}^{N_0} \frac{w_{i_1, i_0} x_{i_0}^\mu}{\sqrt{N_0}} \right) \right]^2 \right\}$$

PARTITION FUNCTION FOR ONE HIDDEN LAYER NN IN THE ASYMPTOTIC LIMIT (1)

$$Z = \int \prod_{i_1}^{N_1} dv_{i_1} \prod_{i_1,i_0}^{N_1,N_0} dw_{i_1 i_0} \, \exp\left\{ -\frac{\lambda_1}{2} \sum_{i_1}^{N_1} v_{i_1}^2 - \frac{\lambda_0}{2} \|w\|^2 - \frac{\beta}{2} \sum_{\mu}^{P} \left[y^\mu - \frac{1}{\sqrt{N_1}} \sum_{i_1}^{N_1} v_{i_1} \sigma \left(\sum_{i_0}^{N_0} \frac{w_{i_1,i_0} x_{i_0}^\mu}{\sqrt{N_0}} \right) \right]^2 \right\}$$

- I want to integrate over the weights of the network (I cannot do it for free)

PARTITION FUNCTION FOR ONE HIDDEN LAYER NN IN THE ASYMPTOTIC LIMIT (1)

$$Z = \int \prod_{i_1}^{N_1} dv_{i_1} \prod_{i_1,i_0}^{N_1,N_0} dw_{i_1 i_0} \, \exp\left\{ -\frac{\lambda_1}{2} \sum_{i_1}^{N_1} v_{i_1}^2 - \frac{\lambda_0}{2} \|w\|^2 - \frac{\beta}{2} \sum_{\mu}^{P} \left[y^\mu - \frac{1}{\sqrt{N_1}} \sum_{i_1}^{N_1} v_{i_1} \sigma\left(\sum_{i_0}^{N_0} \frac{w_{i_1,i_0} x_{i_0}^\mu}{\sqrt{N_0}} \right) \right]^2 \right\}$$

- I want to integrate over the weights of the network (I cannot do it for free)
- I introduce all possible deltas over the pre-activations

$$1 = \int \prod_{\mu}^{P} \prod_{i_1}^{N_1} dh_{i_1}^\mu \, \delta\left(h_{i_1}^\mu - \frac{1}{\sqrt{N_0}} \sum_{i_0}^{N_0} w_{i_1 i_0} x_{i_0}^\mu \right)$$

$$1 = \int \prod_{\mu}^{P} ds^\mu \, \delta\left[s^\mu - \frac{1}{\sqrt{N_1}} \sum_{i}^{N_1} v_{i_1} \sigma(h_{i_1}^\mu) \right]$$

PARTITION FUNCTION FOR ONE HIDDEN LAYER NN IN THE ASYMPTOTIC LIMIT (1)

$$Z = \int \prod_{i_1}^{N_1} dv_{i_1} \prod_{i_1,i_0}^{N_1,N_0} dw_{i_1 i_0} \exp\left\{ -\frac{\lambda_1}{2}\sum_{i_1}^{N_1} v_{i_1}^2 - \frac{\lambda_0}{2}\|w\|^2 - \frac{\beta}{2}\sum_{\mu}^{P} \left[y^\mu - \frac{1}{\sqrt{N_1}}\sum_{i_1}^{N_1} v_{i_1}\sigma\left(\sum_{i_0}^{N_0}\frac{w_{i_1,i_0}x_{i_0}^\mu}{\sqrt{N_0}}\right)\right]^2\right\}$$

- I want to integrate over the weights of the network (I cannot do it for free)
- I introduce all possible deltas over the pre-activations

$$1 = \int \prod_{\mu}^{P} \prod_{i_1}^{N_1} dh_{i_1}^\mu \,\delta\left(h_{i_1}^\mu - \frac{1}{\sqrt{N_0}}\sum_{i_0}^{N_0} w_{i_1 i_0} x_{i_0}^\mu\right)$$

$$1 = \int \prod_{\mu}^{P} ds^\mu \,\delta\left[s^\mu - \frac{1}{\sqrt{N_1}}\sum_{i}^{N_1} v_{i_1}\sigma(h_{i_1}^\mu)\right]$$

Once I employ an integral representation of the deltas I realise all the integrals over the weights are Gaussian

PARTITION FUNCTION FOR ONE HIDDEN LAYER NN IN THE ASYMPTOTIC LIMIT (2): THE CRITICAL STEP

$$Z = \int \prod_{\mu}^{P} \frac{ds^\mu d\bar{s}^\mu}{2\pi} e^{-\frac{\beta}{2}\sum_\mu (y^\mu - s^\mu)^2 + i\sum_\mu^P s^\mu \bar{s}^\mu} \left\{ \int \frac{dq}{\sqrt{2\pi}} e^{-\frac{q^2}{2}} \int d^P h \, P_1(\{h^\mu\}) \delta\left[q - \frac{1}{\sqrt{\lambda_1 N_1}} \sum_\mu \bar{s}^\mu \sigma(h^\mu)\right] \right\}^{N_1}$$

$$P_1(\{h^\mu\}) = \mathcal{N}(0, C) \qquad C_{\mu\nu} = \frac{1}{\lambda_0 N_0} \sum_{i_0}^{N_0} x^\mu_{i_0} x^\nu_{i_0}$$

$$P(q) = \int d^P h \, P_1(\{h^\mu\}) \delta\left[q - \frac{1}{\sqrt{\lambda_1 N_1}} \sum_\mu \bar{s}^\mu \sigma(h^\mu)\right]$$

PARTITION FUNCTION FOR ONE HIDDEN LAYER NN IN THE ASYMPTOTIC LIMIT (2): THE CRITICAL STEP

$$Z = \int \prod_{\mu}^{P} \frac{ds^\mu d\bar{s}^\mu}{2\pi} e^{-\frac{\beta}{2}\sum_\mu (y^\mu - s^\mu)^2 + i\sum_\mu^P s^\mu \bar{s}^\mu} \left\{ \int \frac{dq}{\sqrt{2\pi}} e^{-\frac{q^2}{2}} \int d^P h\, P_1(\{h^\mu\}) \delta\left[q - \frac{1}{\sqrt{\lambda_1 N_1}} \sum_\mu \bar{s}^\mu \sigma(h^\mu) \right] \right\}^{N_1}$$

$$P_1(\{h^\mu\}) = \mathcal{N}(0, C) \qquad C_{\mu\nu} = \frac{1}{\lambda_0 N_0} \sum_{i_0}^{N_0} x_{i_0}^\mu x_{i_0}^\nu$$

$$P(q) = \int d^P h\, P_1(\{h^\mu\}) \delta\left[q - \frac{1}{\sqrt{\lambda_1 N_1}} \sum_\mu \bar{s}^\mu \sigma(h^\mu) \right] \rightarrow \mathcal{N}(0, Q)$$

This probability is Gaussian for the Breuer-Major Theorem (1983)!!

$$Q = \frac{1}{\lambda_1 N_1} \sum_{\mu\nu}^{P} \bar{s}^\mu K_{\mu\nu} \bar{s}^\nu$$

$$K_{\mu\nu}(C) = \int \frac{dt_1 dt_2}{\sqrt{(2\pi)^2 \det \tilde{C}}} e^{-\frac{1}{2}\mathbf{t}^T \tilde{C}^{-1} \mathbf{t}} \sigma(t_1)\sigma(t_2) \qquad \tilde{C} = \begin{pmatrix} C_{\mu\mu} & C_{\mu\nu} \\ C_{\mu\nu} & C_{\nu\nu} \end{pmatrix}$$

This is just the NNGP kernel that describes the infinite-width limit

PARTITION FUNCTION FOR ONE HIDDEN LAYER NN IN THE ASYMPTOTIC LIMIT (3): SADDLE-POINT ACTION

$$Z = \int dQ d\bar{Q} \, \exp\left[-\frac{N_1}{2} S(Q, \bar{Q}) \right]$$

The model is “solved”, in the sense that the partition function is now in a form suitable to saddle-point integration

$$S = -Q\bar{Q} + \log(1+Q) + \frac{\alpha_1}{P} \mathrm{Tr} \log \beta \left[\frac{\mathbb{I}_P}{\beta} + \frac{\bar{Q}K}{\lambda_1} \right] + \frac{\alpha_1}{P} y^\top \left[\frac{\mathbb{I}_P}{\beta} + \frac{\bar{Q}K}{\lambda_1} \right]^{-1} y$$

PARTITION FUNCTION FOR ONE HIDDEN LAYER NN IN THE ASYMPTOTIC LIMIT (3): SADDLE-POINT ACTION

$$Z = \int dQ\, d\bar{Q}\, \exp\left[-\frac{N_1}{2} S(Q,\bar{Q})\right]$$

The model is “solved”, in the sense that the partition function is now in a form suitable to saddle-point integration

$$\boxed{S = -Q\bar{Q} + \log(1+Q) + \frac{\alpha_1}{P}\mathrm{Tr}\log\beta\left[\frac{\mathbb{I}_P}{\beta} + \frac{\bar{Q}K}{\lambda_1}\right] + \frac{\alpha_1}{P} y^\top \left[\frac{\mathbb{I}_P}{\beta} + \frac{\bar{Q}K}{\lambda_1}\right]^{-1} y}$$

explicit formula for the generalisation error

$$\left\langle \epsilon_{\mathrm{g}}(\mathbf{x}^0, y^0) \right\rangle = \left\langle \left(y^0 - f(\mathbf{x}^0)\right)^2 \right\rangle$$

$$= \left[y^0 - \frac{\bar{Q}}{\lambda_1}\sum_{\mu\nu}\kappa_\mu(\mathbf{x}^0)\left(\frac{\mathbb{I}_P}{\beta} + \frac{\bar{Q}K}{\lambda_1}\right)^{-1}_{\mu\nu} y_\nu\right]^2 + \frac{\bar{Q}}{\lambda_1}\left[\kappa_0(\mathbf{x}^0) - \frac{\bar{Q}}{\lambda_1}\sum_{\mu\nu}\kappa_\mu(\mathbf{x}^0)\left(\frac{\mathbb{I}_P}{\beta} + \frac{\bar{Q}K}{\lambda_1}\right)^{-1}_{\mu\nu}\kappa_\nu(\mathbf{x}^0)\right]$$

A CORRESPONDENCE BETWEEN FINITE-WIDTH ONE HIDDEN LAYER ARCHITECTURES AND STUDENT-T PROCESSES

$$p(\{\bar{s}^\mu\}) \sim \left(1 + \frac{1}{\lambda N_1} \sum_{\mu,\nu}^{P} \bar{s}^\mu K_{\mu\nu}(C) \bar{s}^\nu \right)^{-\frac{N_1}{2}} \sim e^{-\frac{1}{2\lambda} \sum_{\mu,\nu}^{P} \bar{s}^\mu K_{\mu\nu}(C) \bar{s}^\nu}$$

$$N_1 \gg P$$

A CORRESPONDENCE BETWEEN FINITE-WIDTH ONE HIDDEN LAYER ARCHITECTURES AND STUDENT-T PROCESSES

$$p(\{\bar{s}^\mu\}) \sim \left(1 + \frac{1}{\lambda N_1} \sum_{\mu,\nu}^{P} \bar{s}^\mu K_{\mu\nu}(C) \bar{s}^\nu\right)^{-\frac{N_1}{2}} \sim e^{-\frac{1}{2\lambda}\sum_{\mu,\nu}^{P} \bar{s}^\mu K_{\mu\nu}(C)\bar{s}^\nu}$$

$$N_1 \gg P$$

This is a **multivariate Student-t distribution!**

A CORRESPONDENCE BETWEEN FINITE-WIDTH ONE HIDDEN LAYER ARCHITECTURES AND STUDENT-T PROCESSES

$$p(\{\bar{s}^\mu\}) \sim \left(1 + \frac{1}{\lambda N_1} \sum_{\mu,\nu}^{P} \bar{s}^\mu K_{\mu\nu}(C) \bar{s}^\nu \right)^{-\frac{N_1}{2}} \sim e^{-\frac{1}{2\lambda}\sum_{\mu,\nu}^{P} \bar{s}^\mu K_{\mu\nu}(C)\bar{s}^\nu}$$

$$N_1 \gg P$$

This is a **multivariate Student-t distribution!**

- **Finite-width one hidden layer neural networks are related to Student-t stochastic processes**

A CORRESPONDENCE BETWEEN FINITE-WIDTH ONE HIDDEN LAYER ARCHITECTURES AND STUDENT-T PROCESSES

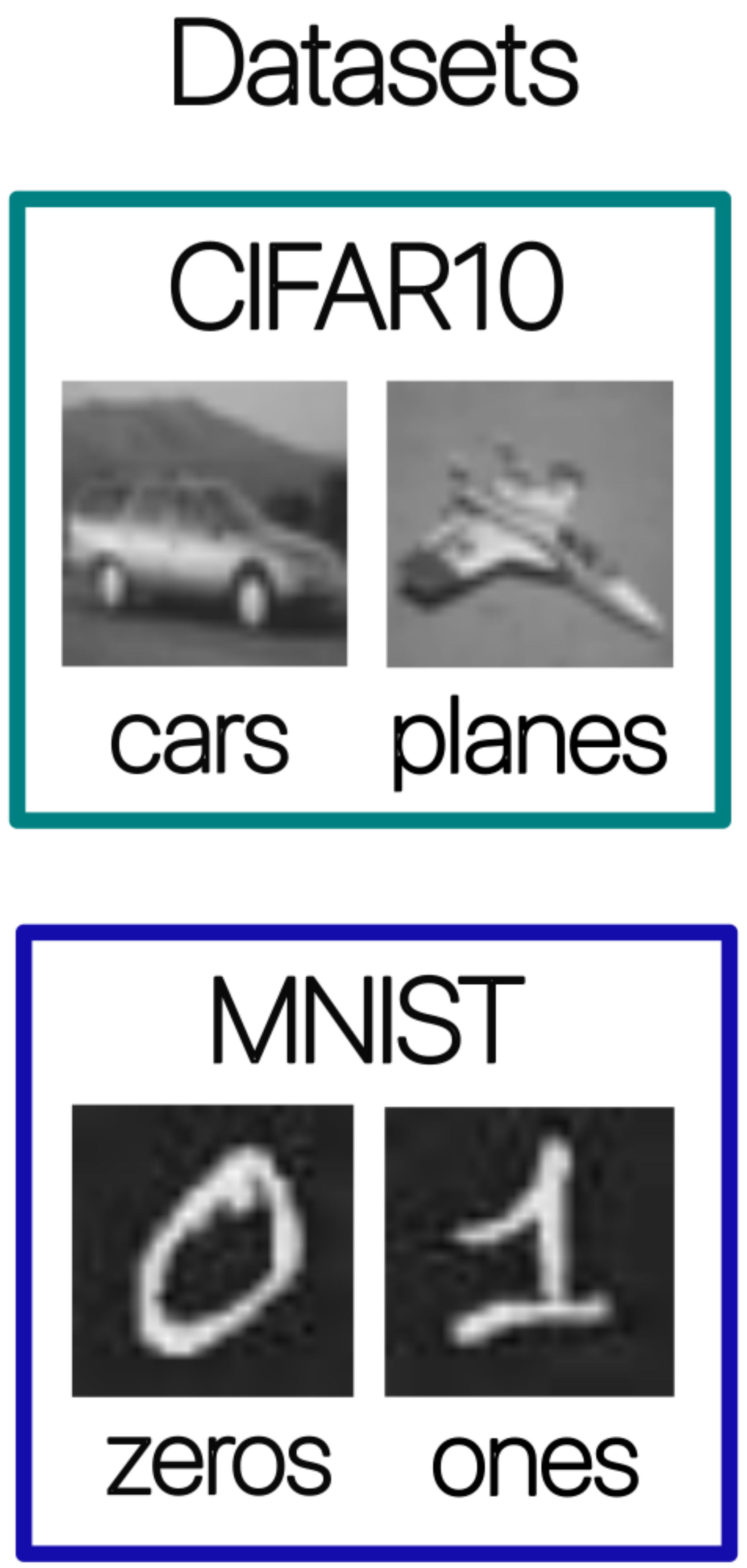
$$p(\{\bar{s}^\mu\}) \sim \left(1 + \frac{1}{\lambda N_1}\sum_{\mu,\nu}^{P} \bar{s}^\mu K_{\mu\nu}(C)\bar{s}^\nu\right)^{-\frac{N_1}{2}} \sim e^{-\frac{1}{2\lambda}\sum_{\mu,\nu}^{P}\bar{s}^\mu K_{\mu\nu}(C)\bar{s}^\nu}$$

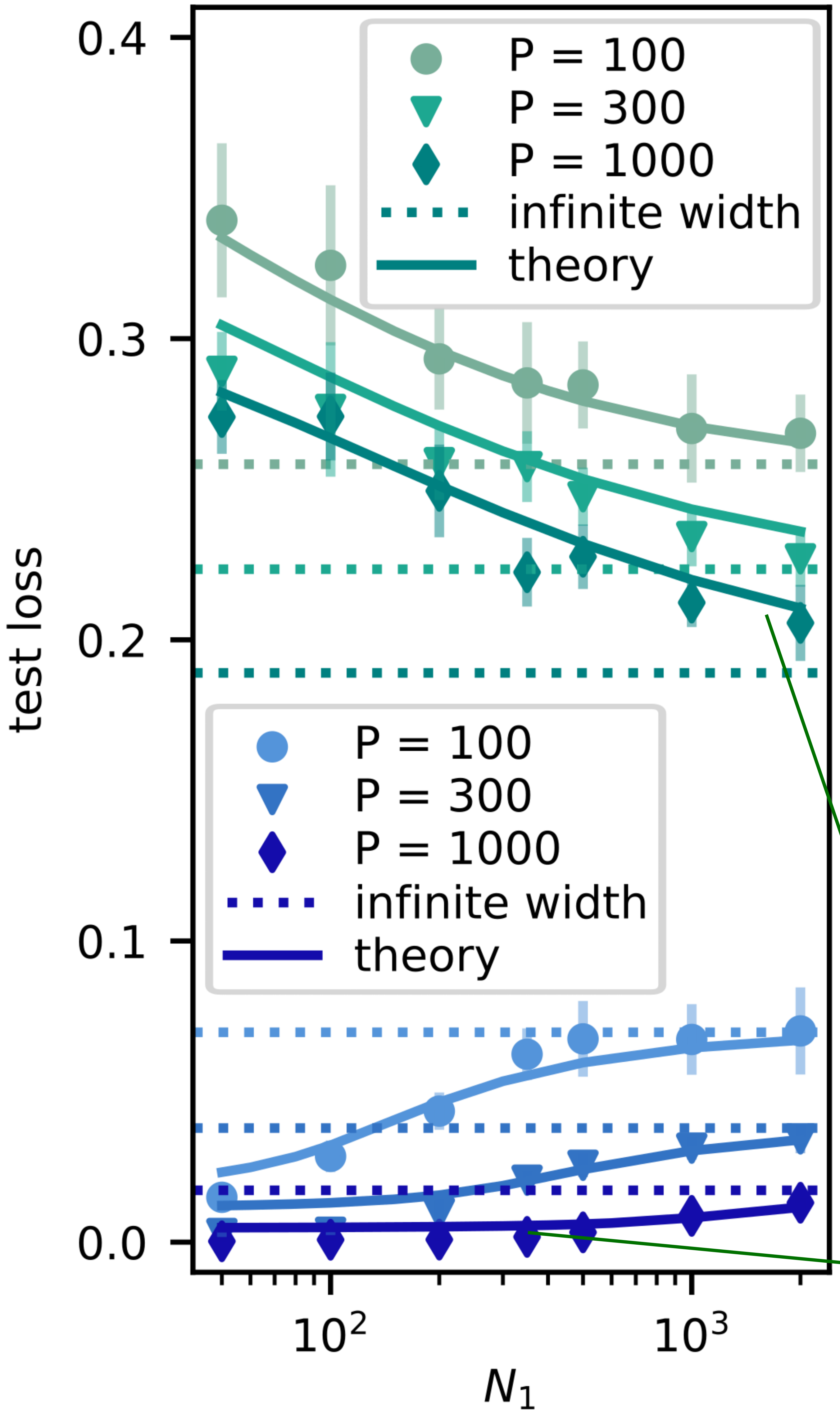
$$N_1 \gg P$$

This is a **multivariate Student-t distribution!**

- **Finite-width one hidden layer neural networks are related to Student-t stochastic processes**
- **Finite-width deep linear networks are also related to Student-t processes!**

VERIFYING THE PREDICTIONS OF THE THEORY AT 1HL USING A DISCRETE LANGEVIN DYNAMICS (1)

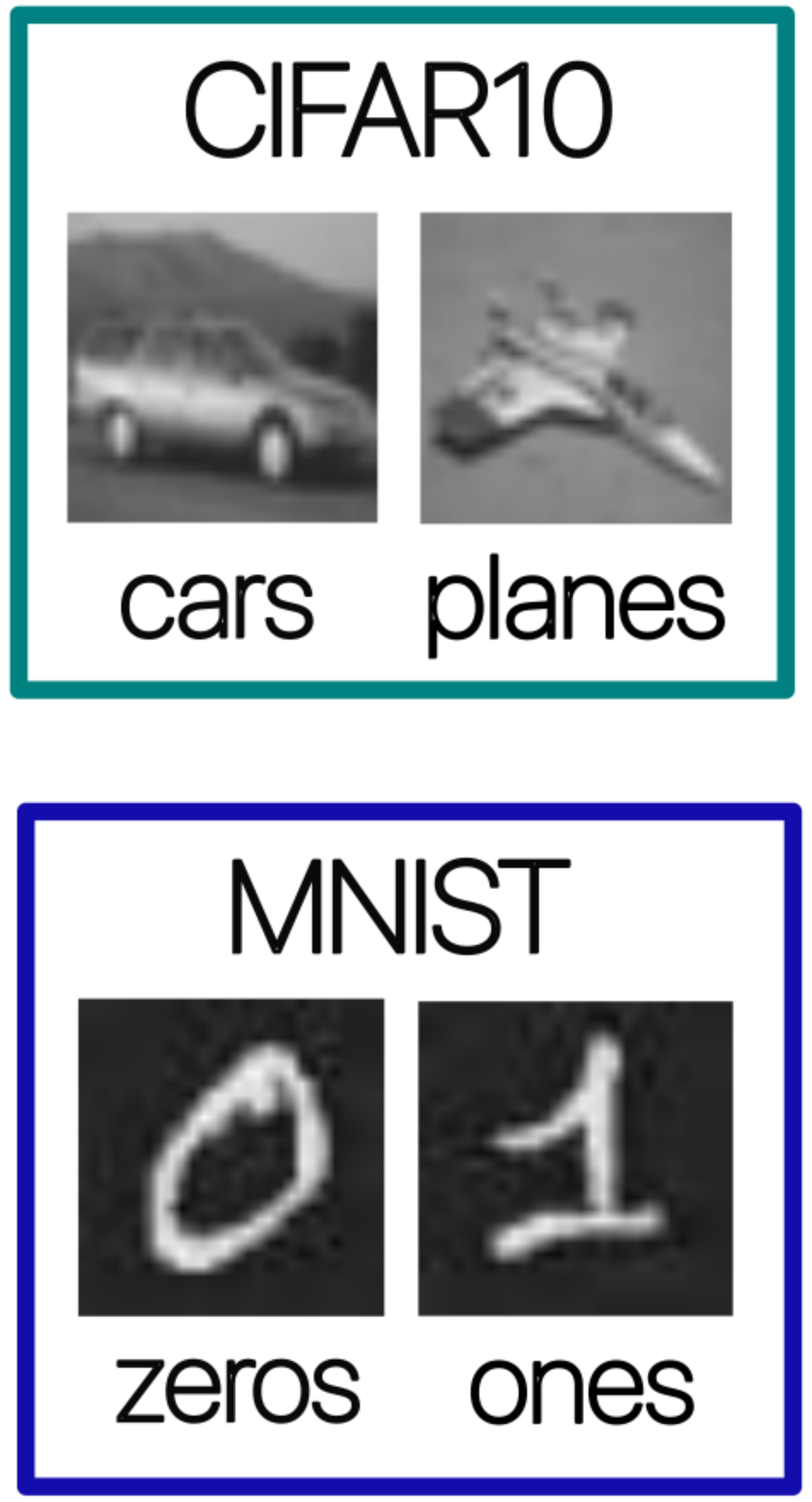


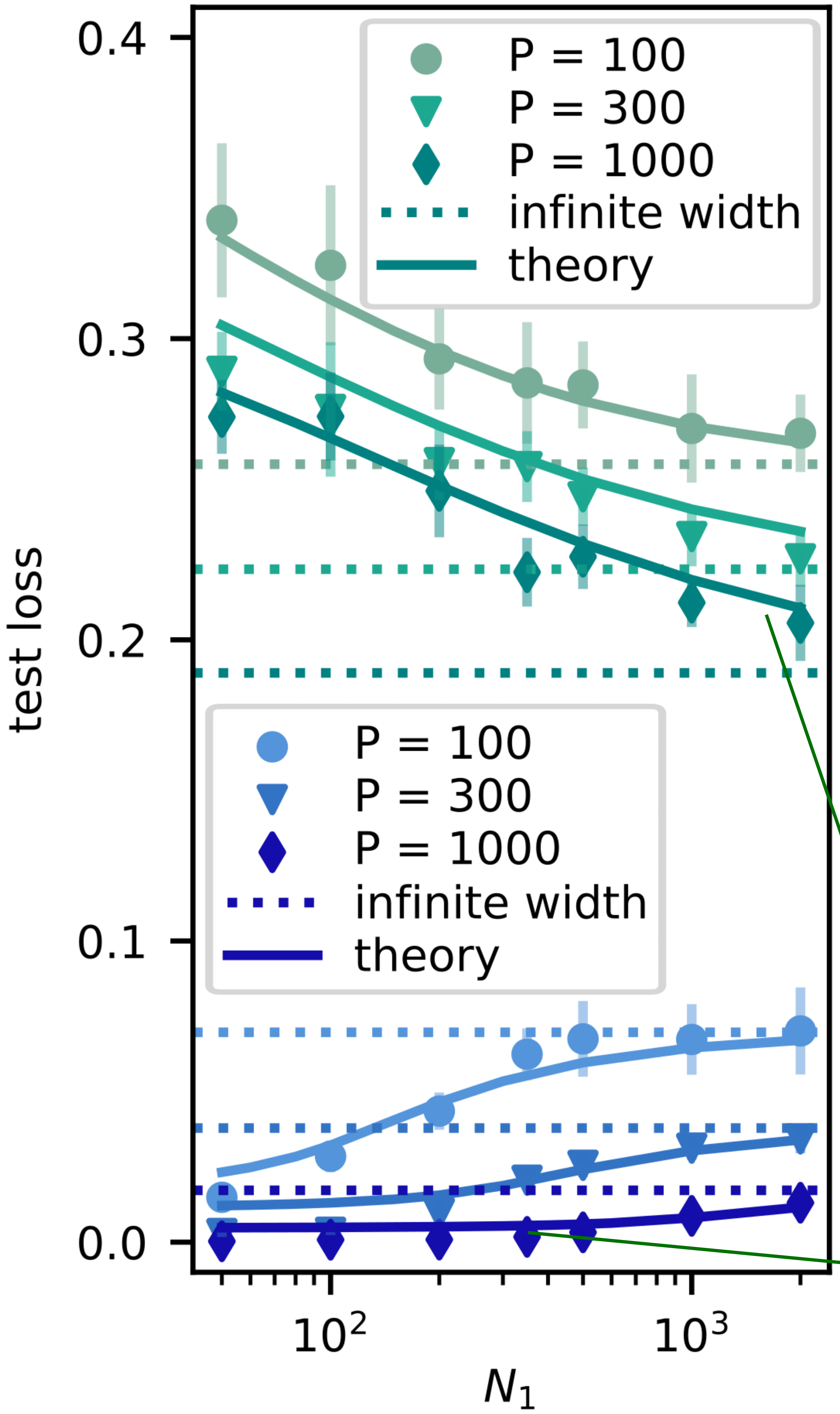


Best test accuracy on the binary classification problem achieved: 86% (CIFAR), 99.9% (MNIST)

VERIFYING THE PREDICTIONS OF THE THEORY AT 1HL USING A DISCRETE LANGEVIN DYNAMICS (1)

Datasets





Learning curves are **monotonically increasing/decreasing** in the range explored (smallest N = 50)

WHY?

analytical criterion

$$y^T K^{-1} y > P$$

ANOTHER CONNECTION WITH STUDENT-T! [Tracey & Wolpert (2018)]

Best test accuracy on the binary classification problem achieved: 86% (CIFAR), 99.9% (MNIST)

VERIFYING THE PREDICTIONS OF THE THEORY AT 1HL USING A DISCRETE LANGEVIN DYNAMICS (2)

Datasets

CIFAR10

cars planes

MNIST

zeros ones

General observations

(i) At T = 0 the bias in constant as a function of N_1 and of the Gaussian prior of the last layer λ_1

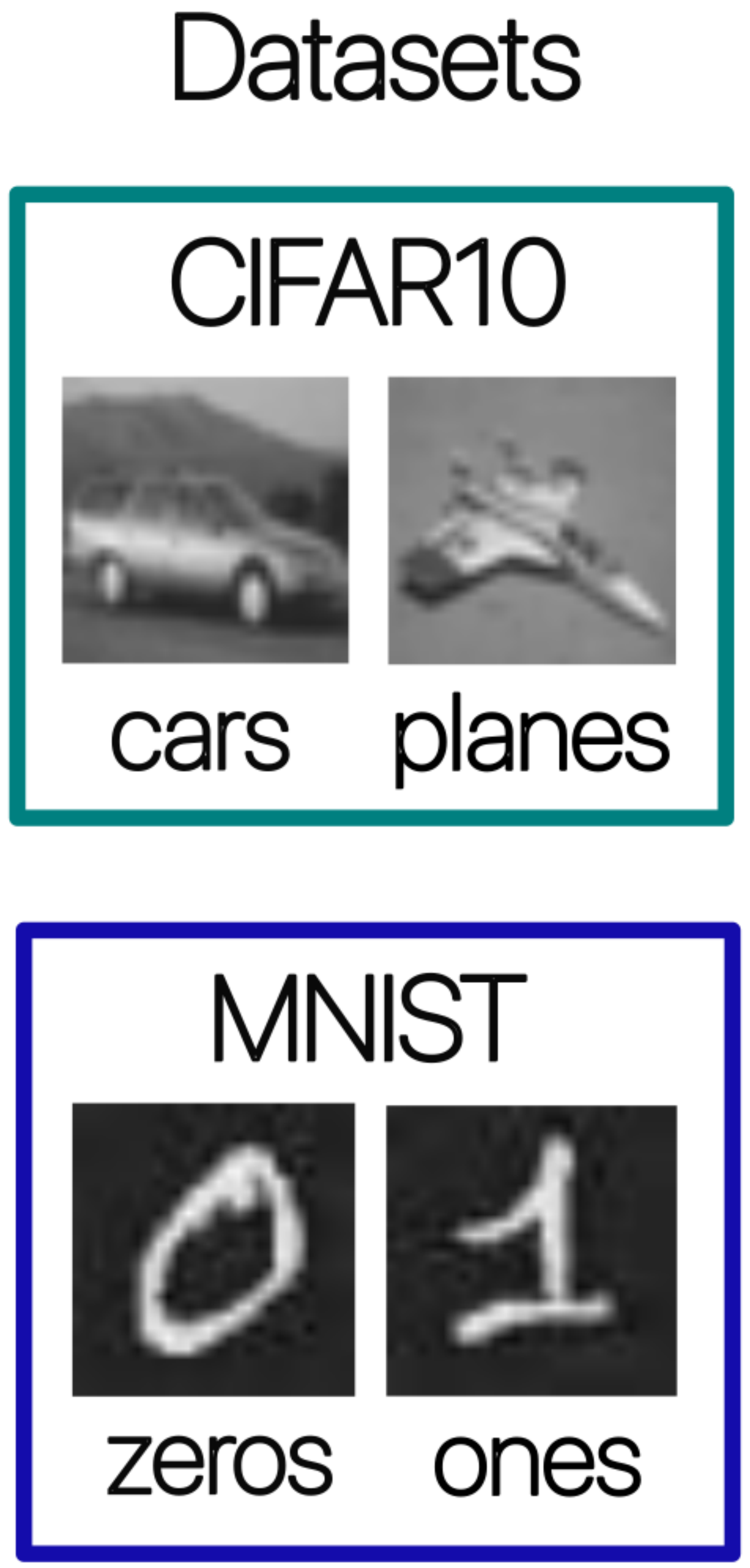
(ii) At T = 0 the variance depends on N_1 and goes to zero as $1/\sqrt{\lambda_1}$

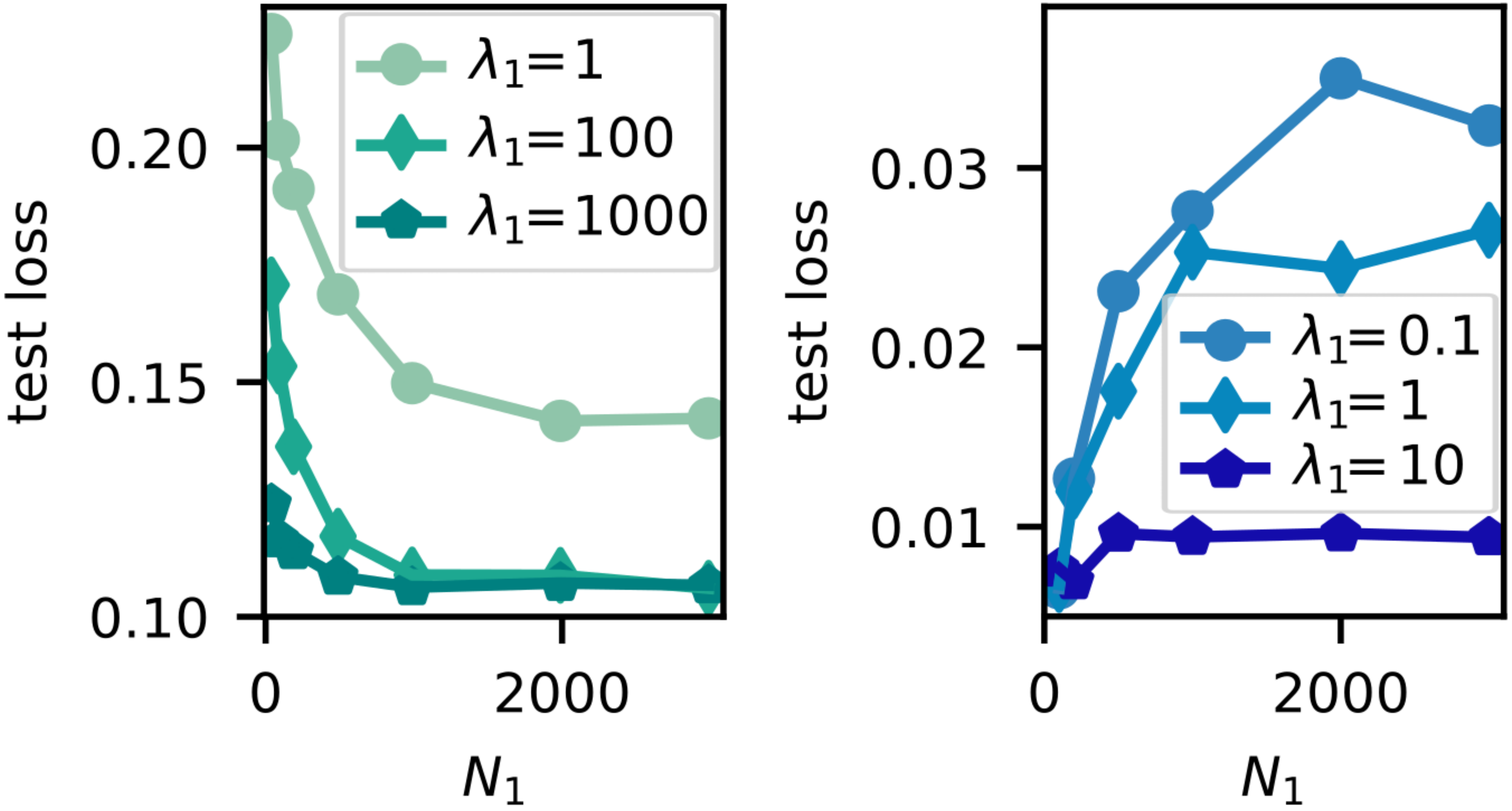
Physical consequences

(i) increasing the magnitude of the last layer Gaussian prior should lead to better generalisation at **ANY** N_1

(ii) For large values of λ_1 the dependence on the size of the hidden layer in the learning curve should disappear

VERIFYING THE PREDICTIONS OF THE THEORY AT 1HL USING A DISCRETE LANGEVIN DYNAMICS (2)





APPROXIMATE PARTITION FUNCTION FOR DNNS WITH ODD ACTIVATION FUNCTION: A RECURRENCE BASED ON STUDENT-T

$$P_{\ell-1}(\{\mathbf{h}^{\mu}_{\ell-1}\}) \longrightarrow P_{\ell}(\{\mathbf{h}^{\mu}_{\ell}\})$$

The goal would be to determine the distribution of the pre-activations at a given layer, given the one at the previous layer

APPROXIMATE PARTITION FUNCTION FOR DNNS WITH ODD ACTIVATION FUNCTION: A RECURRENCE BASED ON STUDENT-T

$$P_{\ell-1}(\{\mathbf{h}^{\mu}_{\ell-1}\}) \longrightarrow P_{\ell}(\{\mathbf{h}^{\mu}_{\ell}\})$$

The goal would be to determine the distribution of the pre-activations at a given layer, given the one at the previous layer

$$\boxed{K^{(R)}_{\ell}(\{\bar{Q}_\ell\}) = \bar{Q}_\ell/\lambda_\ell K \circ \left[K^{(R)}_{\ell-1}(\{\bar{Q}_\ell\})\right] \qquad K^{(R)}_0 = C}$$

APPROXIMATE PARTITION FUNCTION FOR DNNS WITH ODD ACTIVATION FUNCTION: A RECURRENCE BASED ON STUDENT-T

$$P_{\ell-1}(\{\mathbf{h}^{\mu}_{\ell-1}\}) \longrightarrow P_{\ell}(\{\mathbf{h}^{\mu}_{\ell}\})$$

The goal would be to determine the distribution of the pre-activations at a given layer, given the one at the previous layer

$$\boxed{K^{(R)}_{\ell}(\{\bar{Q}_\ell\}) = \bar{Q}_\ell/\lambda_\ell K \circ \left[K^{(R)}_{\ell-1}(\{\bar{Q}_\ell\})\right] \qquad K^{(R)}_0 = C}$$

$$Z_{\mathrm{DNN}} = \int \prod_{\ell=1}^{L} dQ_\ell d\bar{Q}_\ell e^{-\frac{N_L}{2} S_{\mathrm{DNN}}(\{Q_\ell, \bar{Q}_\ell\})}$$

APPROXIMATE PARTITION FUNCTION FOR DNNS WITH ODD ACTIVATION FUNCTION: A RECURRENCE BASED ON STUDENT-T

$$P_{\ell-1}(\{\mathbf{h}^\mu_{\ell-1}\}) \longrightarrow P_\ell(\{\mathbf{h}^\mu_\ell\})$$

The goal would be to determine the distribution of the pre-activations at a given layer, given the one at the previous layer

$$K_\ell^{(R)}(\{\bar{Q}_\ell\}) = \bar{Q}_\ell/\lambda_\ell K \circ \left[K_{\ell-1}^{(R)}(\{\bar{Q}_\ell\})\right] \qquad K_0^{(R)} = C$$

$$Z_{\mathrm{DNN}} = \int \prod_{\ell=1}^{L} dQ_\ell d\bar{Q}_\ell e^{-\frac{N_L}{2} S_{\mathrm{DNN}}(\{Q_\ell,\bar{Q}_\ell\})}$$

Effective action for finite-width fully-connected architectures with L hidden layers

$$S_{\mathrm{DNN}} = \sum_{\ell=1}^{L} \frac{\alpha_L}{\alpha_\ell}\left[-Q_\ell\bar{Q}_\ell + \log(1+Q_\ell)\right] + \frac{\alpha_L}{P}\operatorname{Tr}\log\beta\left(\frac{\mathbb{I}_P}{\beta} + K_L^{(R)}(\{\bar{Q}_\ell\})\right)$$
$$+ \frac{\alpha_L}{P} y^T \left(\frac{\mathbb{I}_P}{\beta} + K_L^{(R)}(\{\bar{Q}_\ell\})\right)^{-1} y$$

APPROXIMATE PARTITION FUNCTION FOR DEEP NEURAL NETWORKS: A RECURRENCE BASED ON STUDENT-T

IMPORTANT!

From this effective theory I am able to recover the Li-Sompolinsky heuristic theory valid for ReLU activation found in the isotropic limit

$$Z_{\mathrm{DNN}} = \int \prod_{\ell=1}^{L} dQ_\ell d\bar{Q}_\ell e^{-\frac{N_L}{2} S_{\mathrm{DNN}}(\{Q_\ell,\bar{Q}_\ell\})}$$

Effective action for finite-width fully-connected architectures with L hidden layers

$$S_{\mathrm{DNN}} = \sum_{\ell=1}^{L} \frac{\alpha_L}{\alpha_\ell}\left[-Q_\ell \bar{Q}_\ell + \log(1+Q_\ell)\right] + \frac{\alpha_L}{P}\mathrm{Tr}\log\beta\left(\frac{\mathbb{I}_P}{\beta} + K_L^{(R)}(\{\bar{Q}_\ell\})\right) + \frac{\alpha_L}{P} y^T \left(\frac{\mathbb{I}_P}{\beta} + K_L^{(R)}(\{\bar{Q}_\ell\})\right)^{-1} y$$

GENERALISING THE APPROACH TO NON-ODD ACTIVATION FUNCTION: BEYOND LI-SOMPOLINSKY HEURISTIC THEORY

Let us go back to the derivation of the one hidden layer effective action…

$$P(q) = \int d^P h \, P_1(\{h^\mu\}) \delta\left[q - \frac{1}{\sqrt{\lambda N_1}} \sum_\mu \bar{s}^\mu \sigma\left(h^\mu\right) \right] \to \mathcal{N}(0, Q)$$

This random variable has zero mean **ONLY** if the activation function is odd

GENERALISING THE APPROACH TO NON-ODD ACTIVATION FUNCTION: BEYOND LI-SOMPOLINSKY HEURISTIC THEORY

Let us go back to the derivation of the one hidden layer effective action…

$$P(q) = \int d^P h \, P_1(\{h^\mu\}) \delta\left[q - \frac{1}{\sqrt{\lambda N_1}} \sum_\mu \bar{s}^\mu \sigma\left(h^\mu\right)\right] \to \mathcal{N}(0, Q)$$

This random variable has zero mean **ONLY** if the activation function is odd

$$\langle q \rangle = \frac{1}{\sqrt{N_1 \lambda}} \sum_{\mu=1}^{P} \bar{s}^\mu m^\mu \qquad m^\mu = \int \frac{dt}{\sqrt{2\pi C_{\mu\mu}}} e^{-t^2/(2C_{\mu\mu})} \sigma(t)$$

GENERALISING THE APPROACH TO NON-ODD ACTIVATION FUNCTION: BEYOND LI-SOMPOLINSKY HEURISTIC THEORY

Let us go back to the derivation of the one hidden layer effective action…

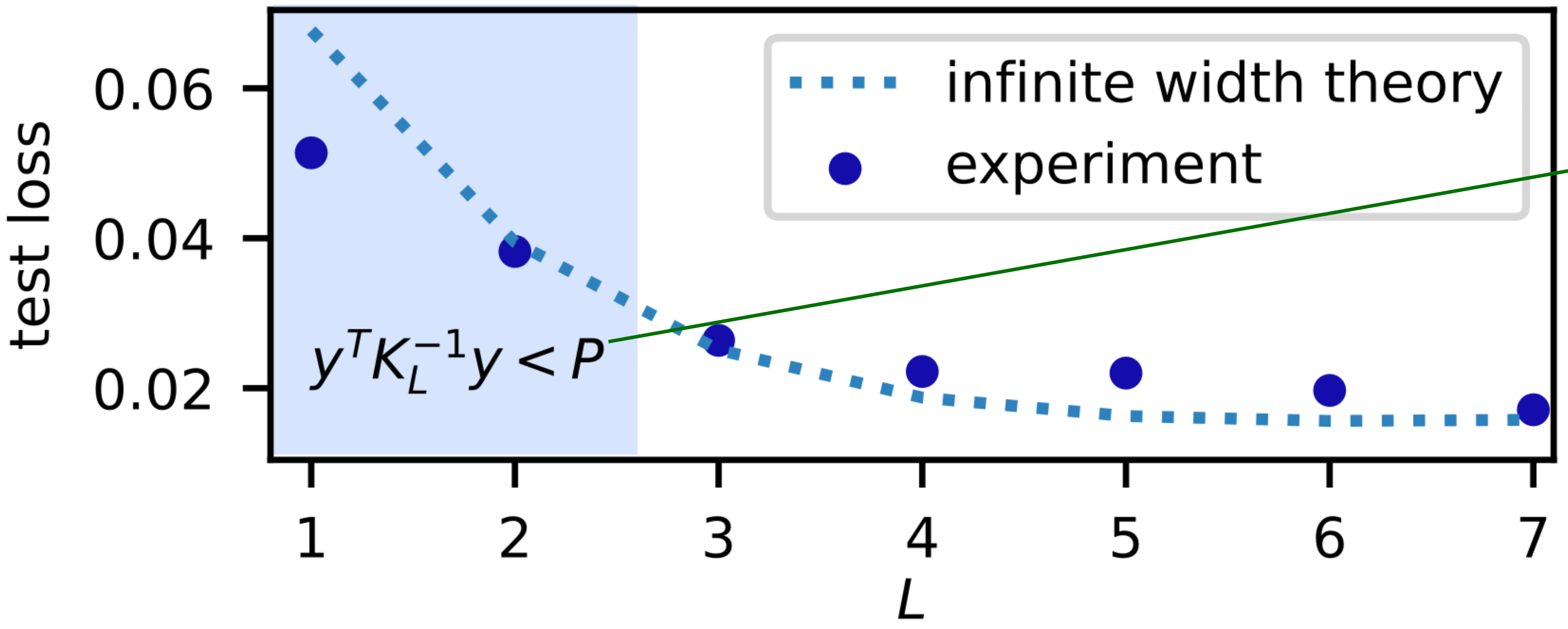
$$P(q) = \int d^P h \, P_1(\{h^\mu\})\delta\left[q - \frac{1}{\sqrt{\lambda N_1}}\sum_\mu \bar{s}^\mu \sigma\left(h^\mu\right)\right] \to \mathcal{N}(0, Q)$$

This random variable has zero mean **ONLY** if the activation function is odd

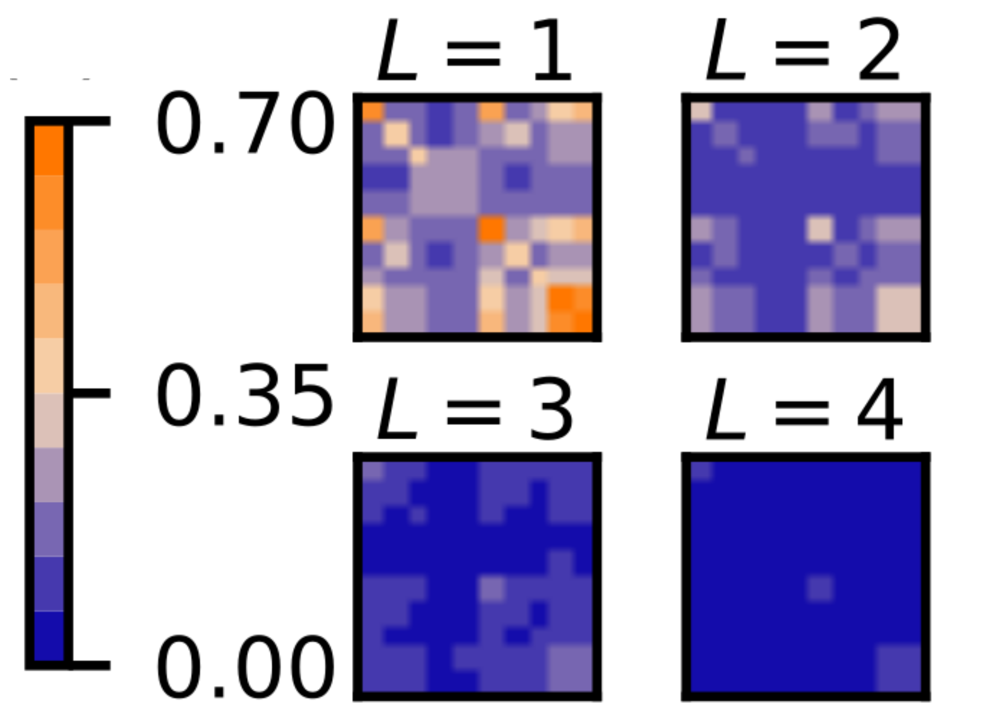
$$\langle q \rangle = \frac{1}{\sqrt{N_1 \lambda}}\sum_{\mu=1}^{P} \bar{s}^\mu m^\mu \qquad m^\mu = \int \frac{dt}{\sqrt{2\pi C_{\mu\mu}}} e^{-t^2/(2C_{\mu\mu})}\sigma(t)$$

$$\boxed{\bar{Q}K \to \bar{Q}K - \left(\bar{Q} + \frac{1}{1+Q}\right)K^{(1)} \qquad K^{(1)}_{\mu\nu} = m^\mu m^\nu}$$

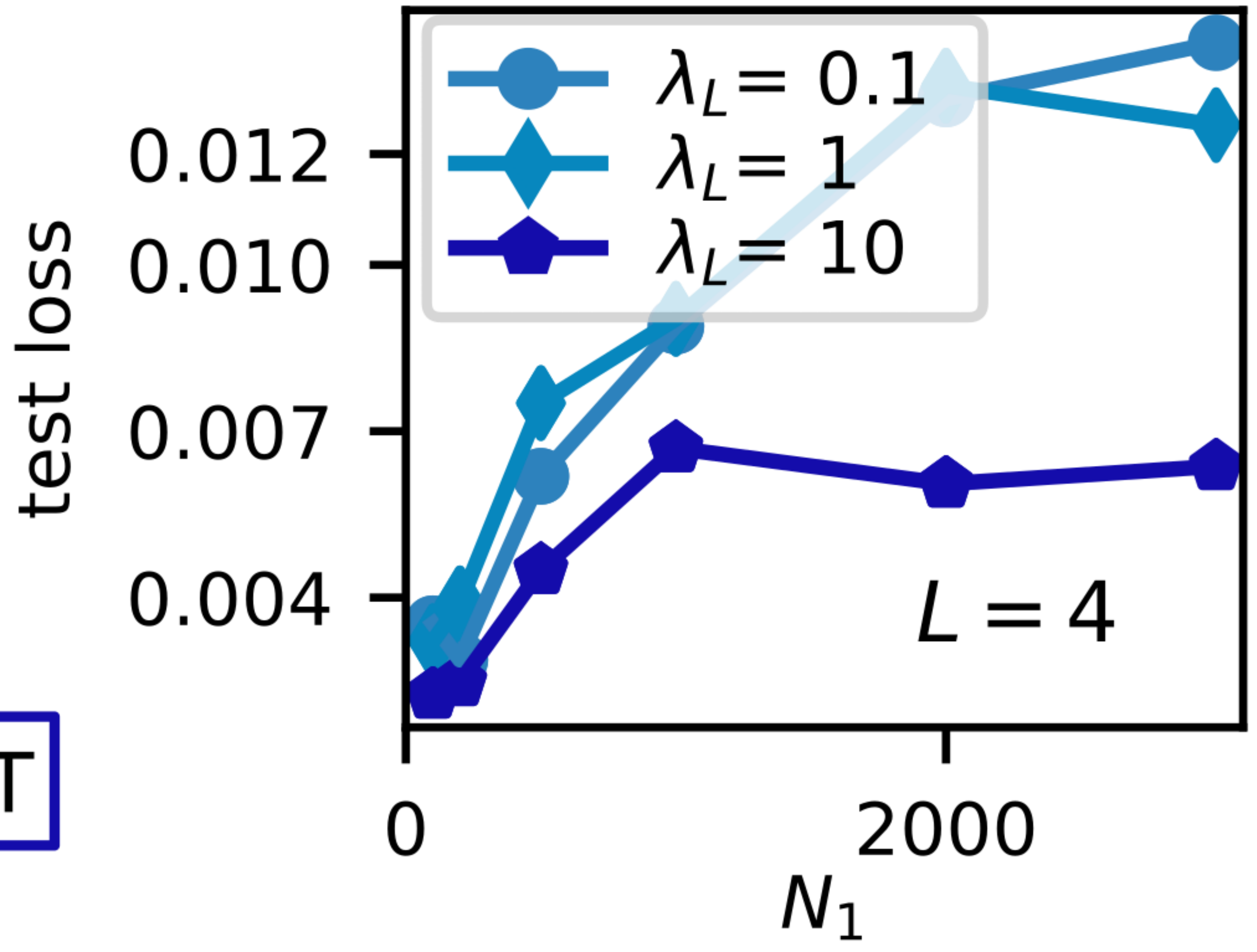
PRELIMINARY VERIFICATION OF THE THEORY AT L LAYERS



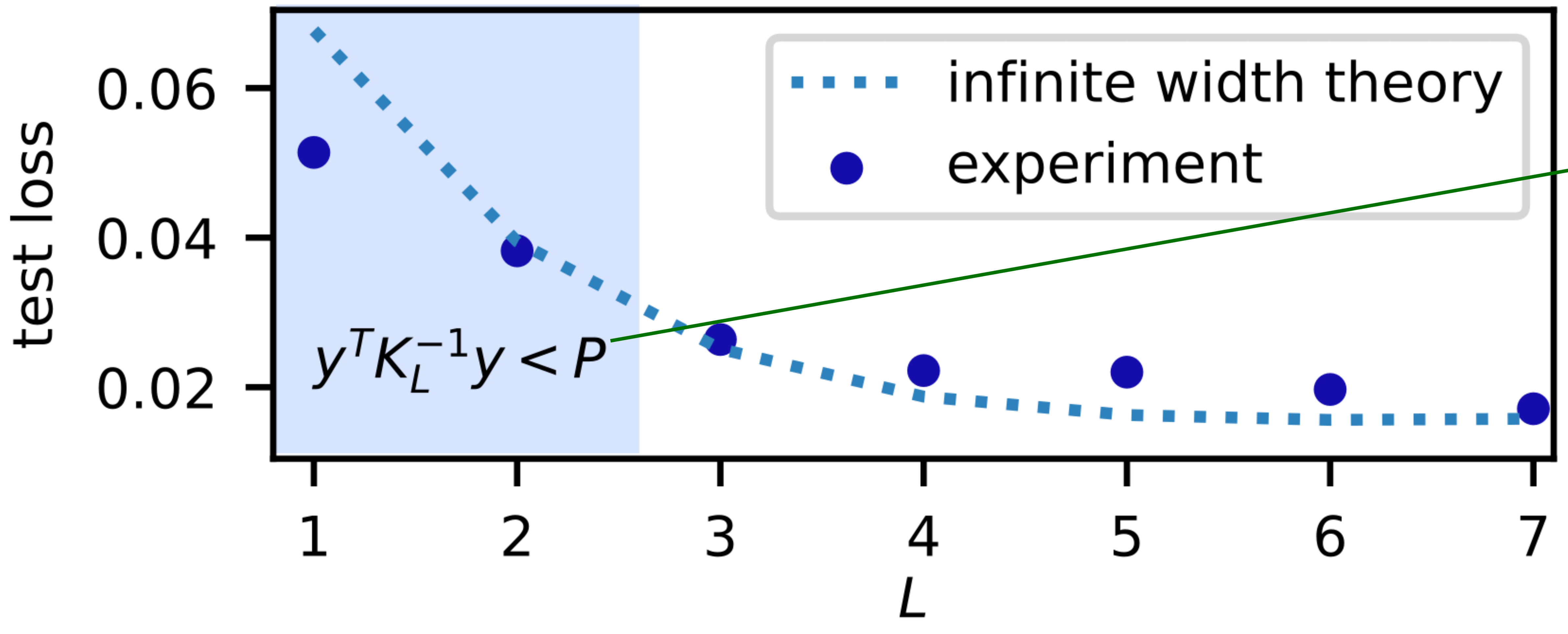
a criterion to establish if finite-width networks will outperform their infinite-width counterpart holds for ReLU



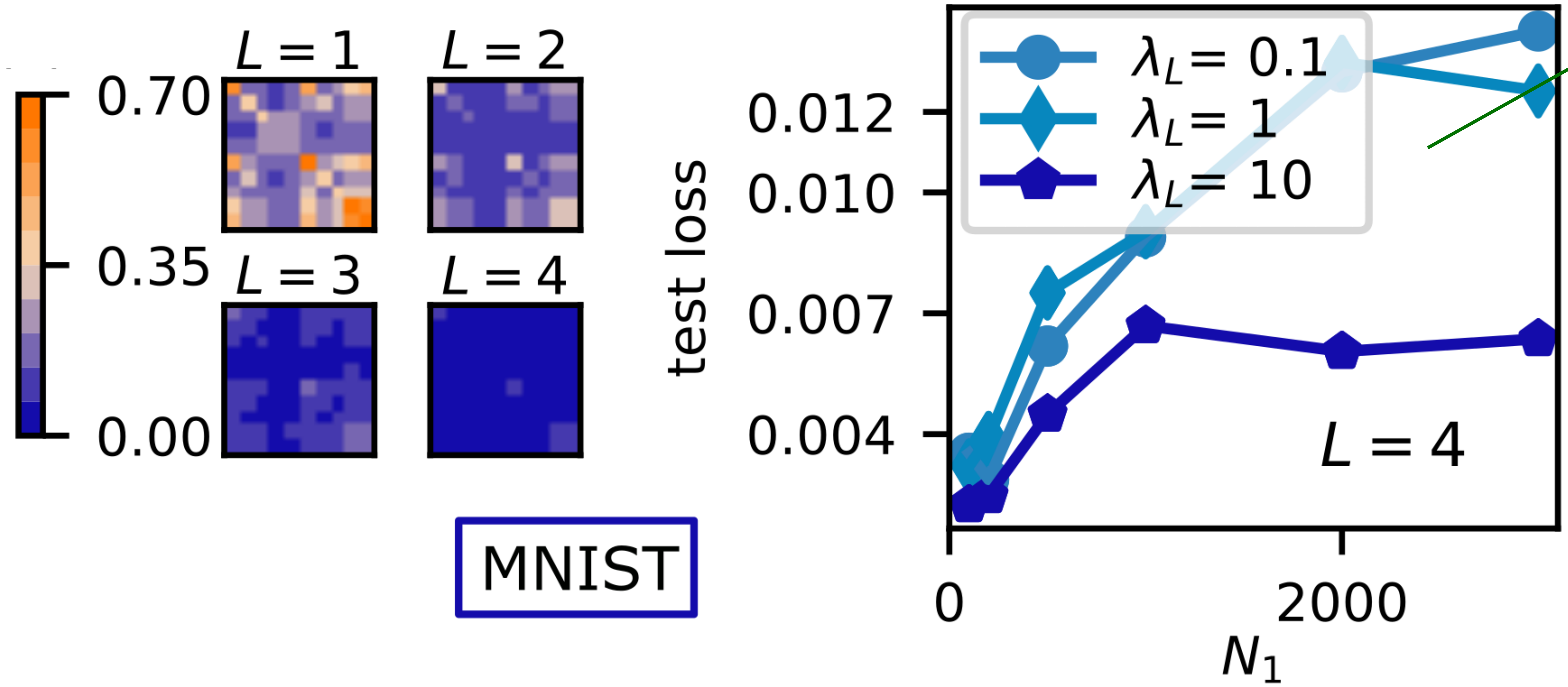
MNIST



PRELIMINARY VERIFICATION OF THE THEORY AT L LAYERS

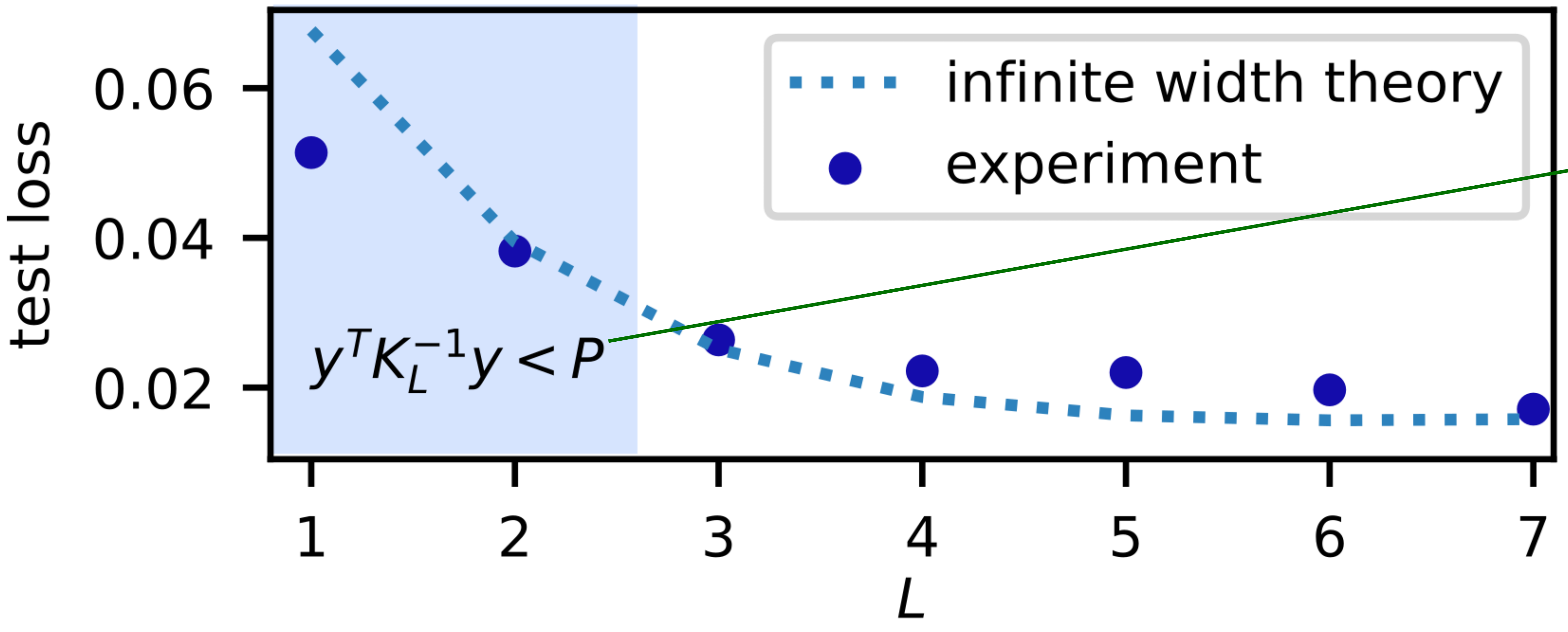


a criterion to establish if finite-width networks will outperform their infinite-width counterpart holds for ReLU

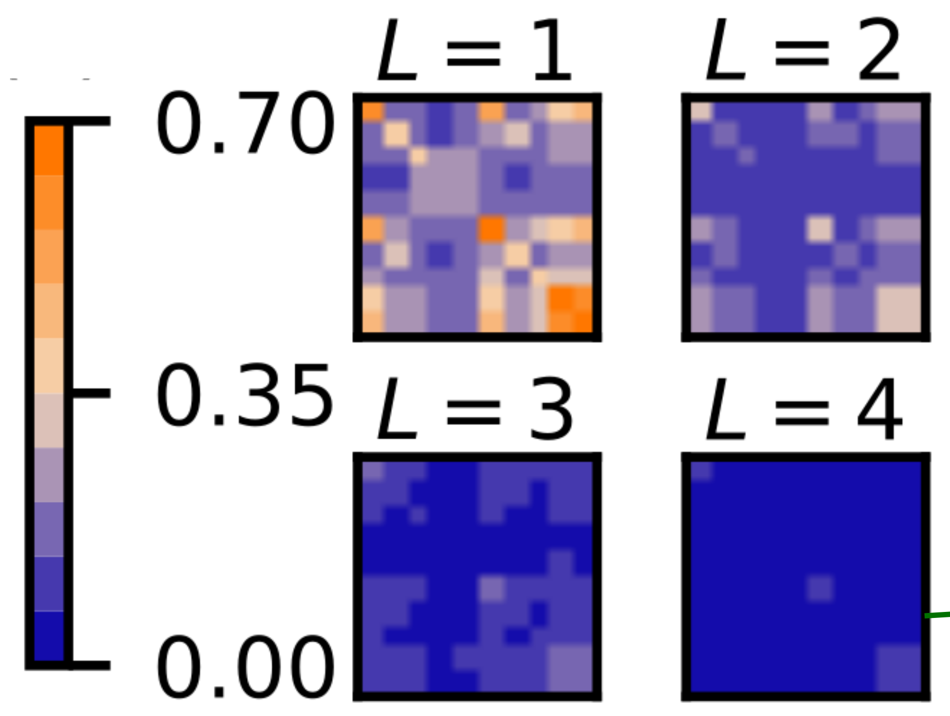


The same reasoning on the Gaussian prior of the last layer holds, but **for L layer the bias is not constant as a function of N!**

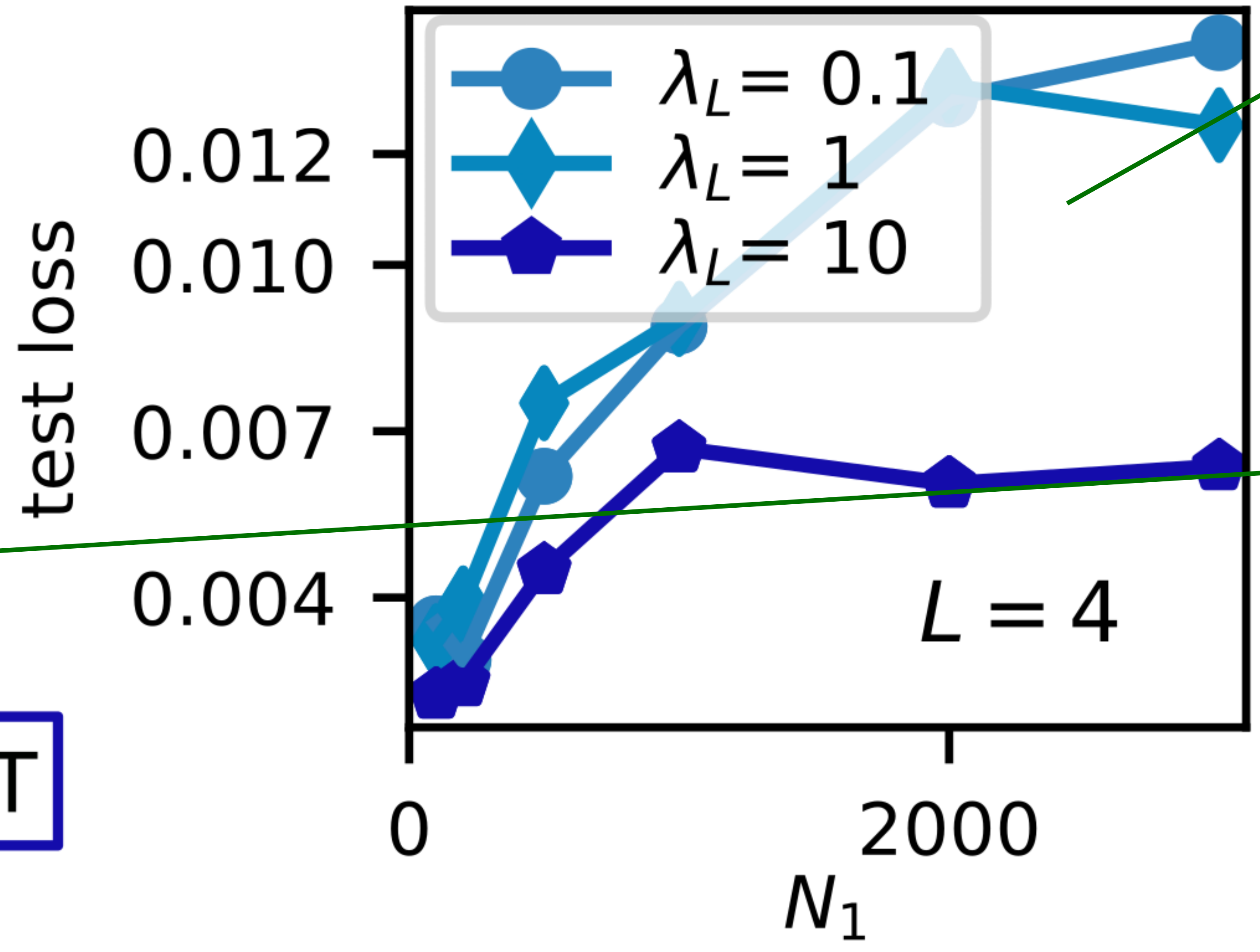
PRELIMINARY VERIFICATION OF THE THEORY AT L LAYERS



a criterion to establish if finite-width networks will outperform their infinite-width counterpart holds for ReLU



MNIST



The same reasoning on the Gaussian prior of the last layer holds, but **for L layer the bias is not constant as a function of N!**

For ReLU, after a certain critical L, infinite-width outperforms finite width, since the NNGP kernel develops at least one almost singular eigenvalue

CONCLUSIONS AND FUTURE PERSPECTIVES

CONCLUSIONS AND FUTURE PERSPECTIVES

- An approach to investigate the statistical physics of neural networks with L fully-connected hidden layers beyond the infinite-width limit

CONCLUSIONS AND FUTURE PERSPECTIVES

- An approach to investigate the statistical physics of neural networks with L fully-connected hidden layers beyond the infinite-width limit
- An intriguing connection between finite-width deep neural networks and Student-t stochastic processes

$$P\,, N_\ell \to \infty \qquad \alpha_\ell = \frac{P}{N_\ell}$$

CONCLUSIONS AND FUTURE PERSPECTIVES

- An approach to investigate the statistical physics of neural networks with L fully-connected hidden layers beyond the infinite-width limit
- An intriguing connection between finite-width deep neural networks and Student-t stochastic processes

$$P\,, N_\ell \to \infty \qquad \alpha_\ell = \frac{P}{N_\ell}$$

- Effective theory for gradient descent dynamics? Convolutions? Feature learning? Generalisation performance at finite-width and edge of chaos? Role of skip connections (residual networks)?

THANKS!

Sebastiano Ariosto

Mauro Pastore

Francesco Ginelli

Marco Gherardi

Rosalba Pacelli

[arXiv:2209.04882 (2022)]