The input-output Jacobian and initialization of neural networks - our contribution for ResNets and some earlier results

Wojciech Tarnowski, Piotr Warchoł, Stanisław Jastrzębski,
Jacek Tabor, Maciej Nowak

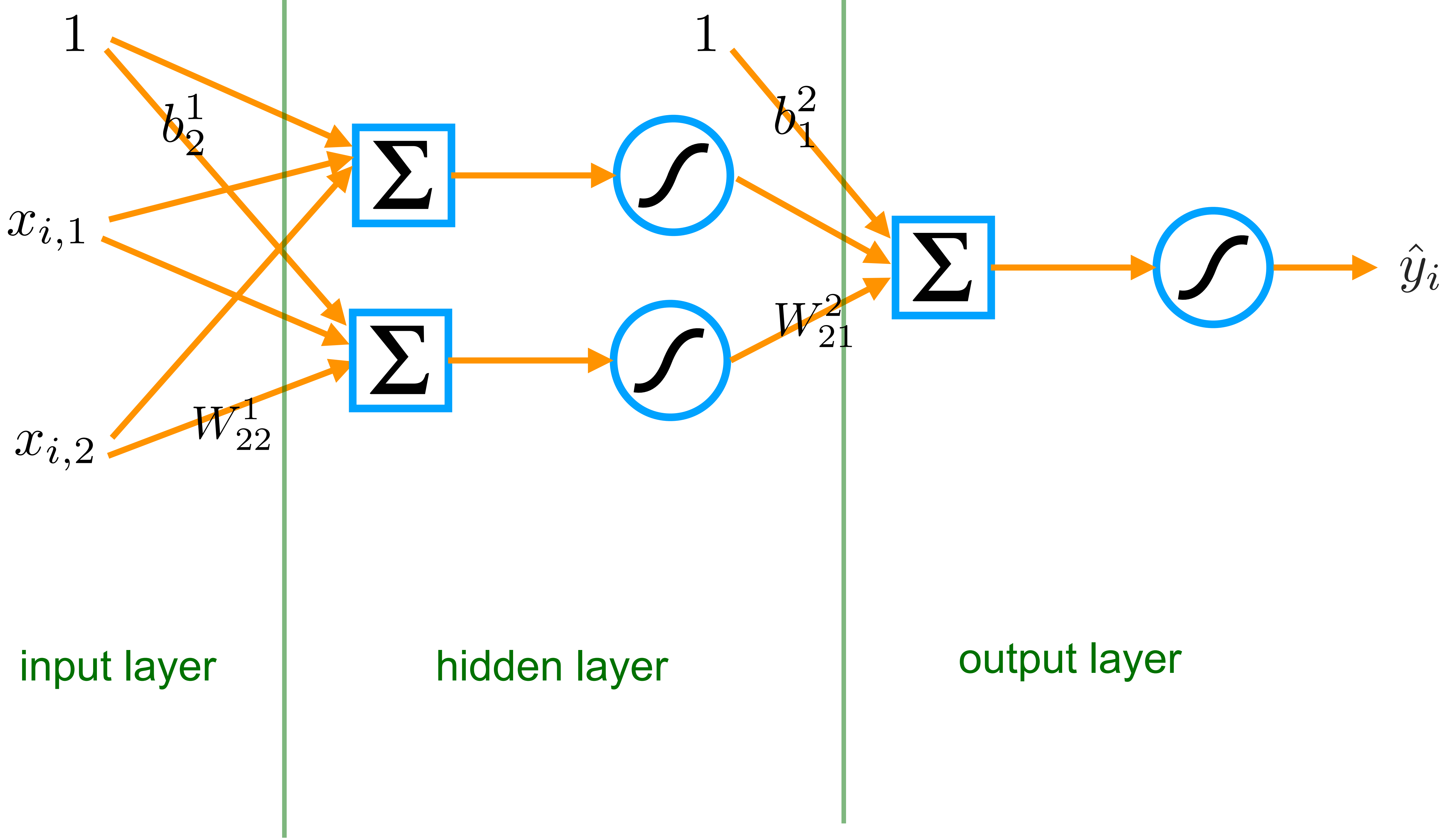
JAGIELLONIAN UNIVERSITY
IN KRAKÓW

We tackle the problem of initialization of deep Residual Neural Networks with Random Matrix and Free Probability Theories.

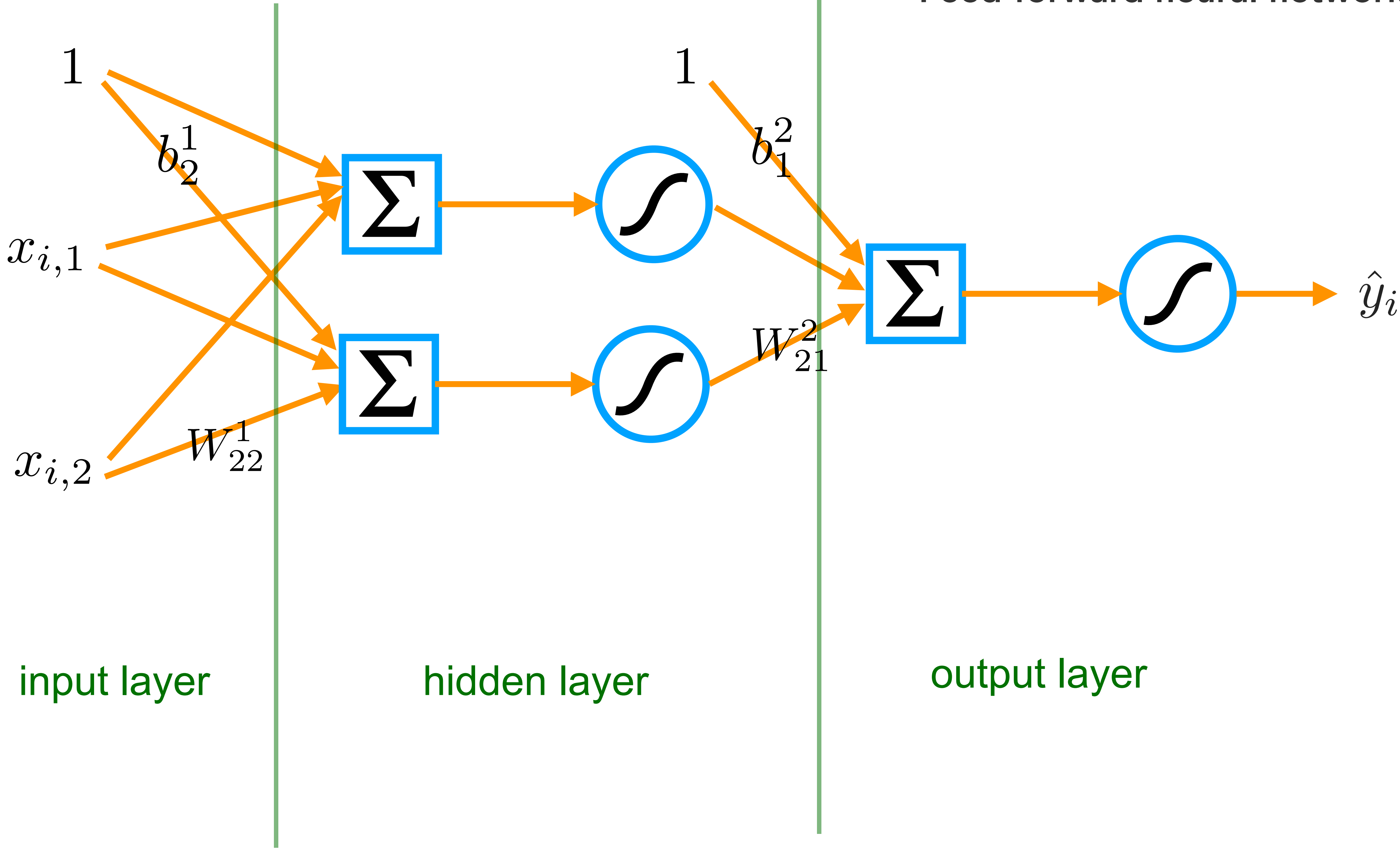
This is done by making sure the spectrum of the input-output Jacobian is concentrated around one. This is called dynamical isometry.

$$J_{ik} = \frac{\partial x_i^L}{\partial x_k^0}$$

Feed-forward neural network

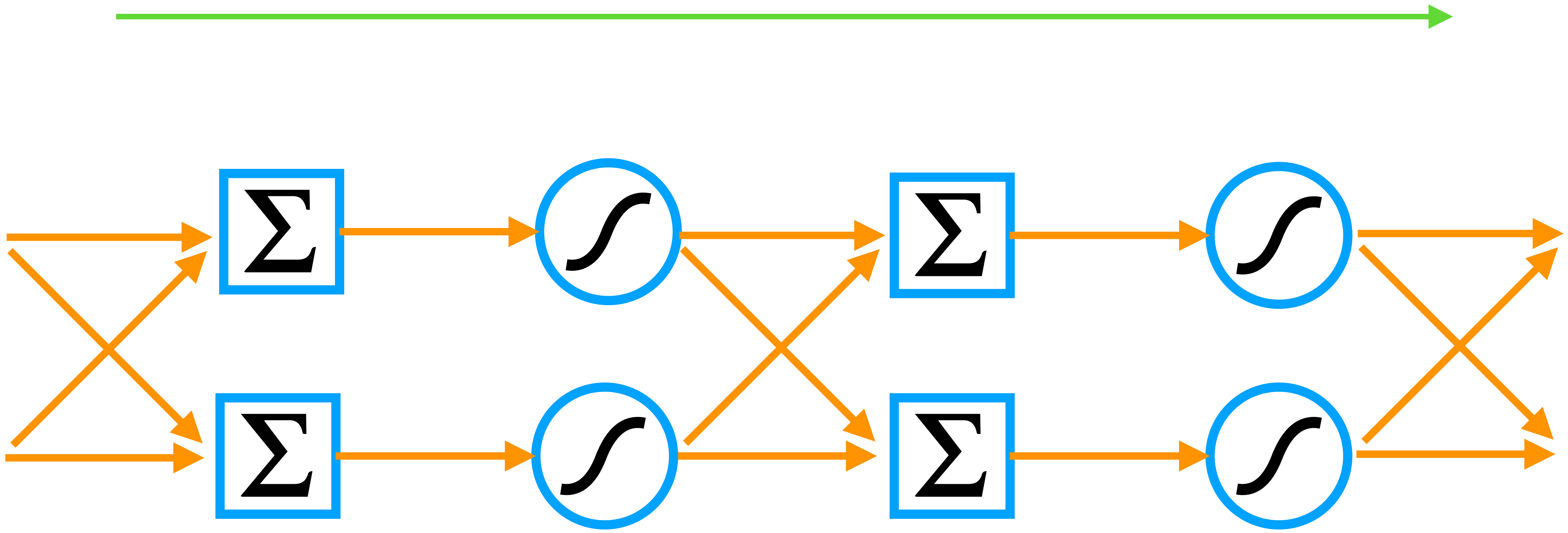


Feed-forward neural network



Signal propagation: $$\mathbf{x^l} = \phi(\mathbf{h^l}), \quad \mathbf{h^l} = \mathbf{W^l x^{l-1}} + \mathbf{b^l}$$

L - number of layers in the network



Signal propagation:

$$\mathbf{x^l} = \phi(\mathbf{h^l}), \quad \mathbf{h^l} = \mathbf{W^l x^{l-1}} + \mathbf{b^l}$$

Signal propagation: $$\mathbf{x}^{\mathbf{l}} = \phi(\mathbf{h}^{\mathbf{l}}), \quad \mathbf{h}^{\mathbf{l}} = \mathbf{W}^{\mathbf{l}}\mathbf{x}^{\mathbf{l}-\mathbf{1}} + \mathbf{b}^{\mathbf{l}}$$

elements of weight matrices and bias vectors - i.i.d. Gaussian with mean 0 and variances: σ_w^2/N_{l-1} and σ_b^2

This constitutes a maximal entropy distribution over ensembles of neural networks under some conditions on the first two moments

Idea 1 (Poole et al. arXiv:1606.05340):
Consider a single input $\mathbf{x}^0$
How will it change while propagating through the network?

We track its length defined by $$q^l = \frac{1}{N_l}\sum_{i=1}^{N_l}(\mathbf{h}_i^l)^2$$

$$q^l = \frac{1}{N_l} \sum_{i=1}^{N_l} (\mathbf{h}_i^l)^2$$

But this is the second moment of the empirical distribution of the pre-activations

Signal propagation: $\mathbf{x}^\mathbf{l} = \phi(\mathbf{h}^\mathbf{l}), \quad \mathbf{h}^\mathbf{l} = \mathbf{W}^\mathbf{l}\mathbf{x}^{\mathbf{l}-\mathbf{1}} + \mathbf{b}^\mathbf{l}$

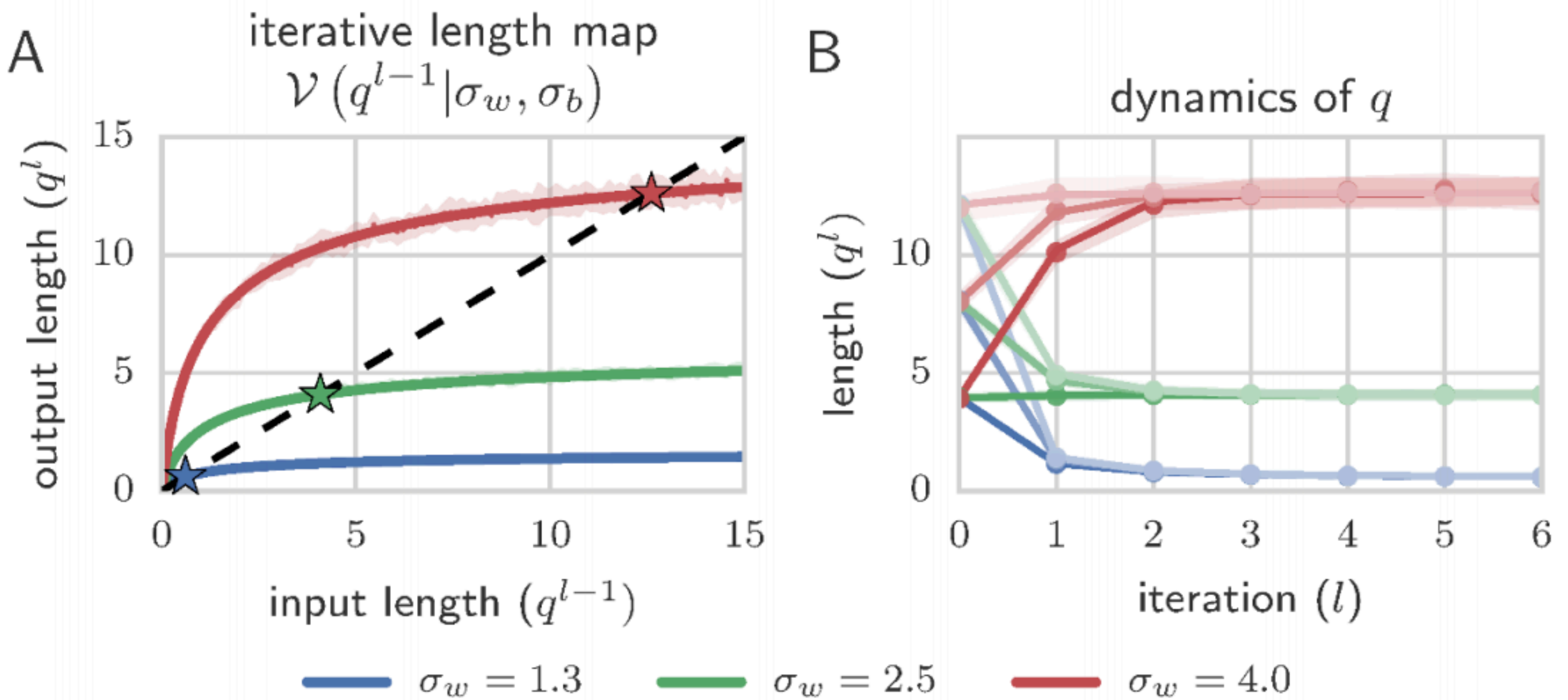
Central limit theorem plus ergodicity arguments:

$$q^l = \mathcal{V}(q^{l-1} \,|\, \sigma_w, \sigma_b) \equiv \sigma_w^2 \int \mathcal{D}z \, \phi\left(\sqrt{q^{l-1}} z\right)^2 + \sigma_b^2, \quad \text{for} \quad l = 2, \ldots, D$$

$$\mathcal{D}z = \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad q^1 = \sigma_w^2 q^0 + \sigma_b^2 \qquad q^0 = \frac{1}{N_0} \mathbf{x}^0 \cdot \mathbf{x}^0$$

An iterative map describing the propagation of the variance of the (Gaussian) probability distribution of the pre-activations

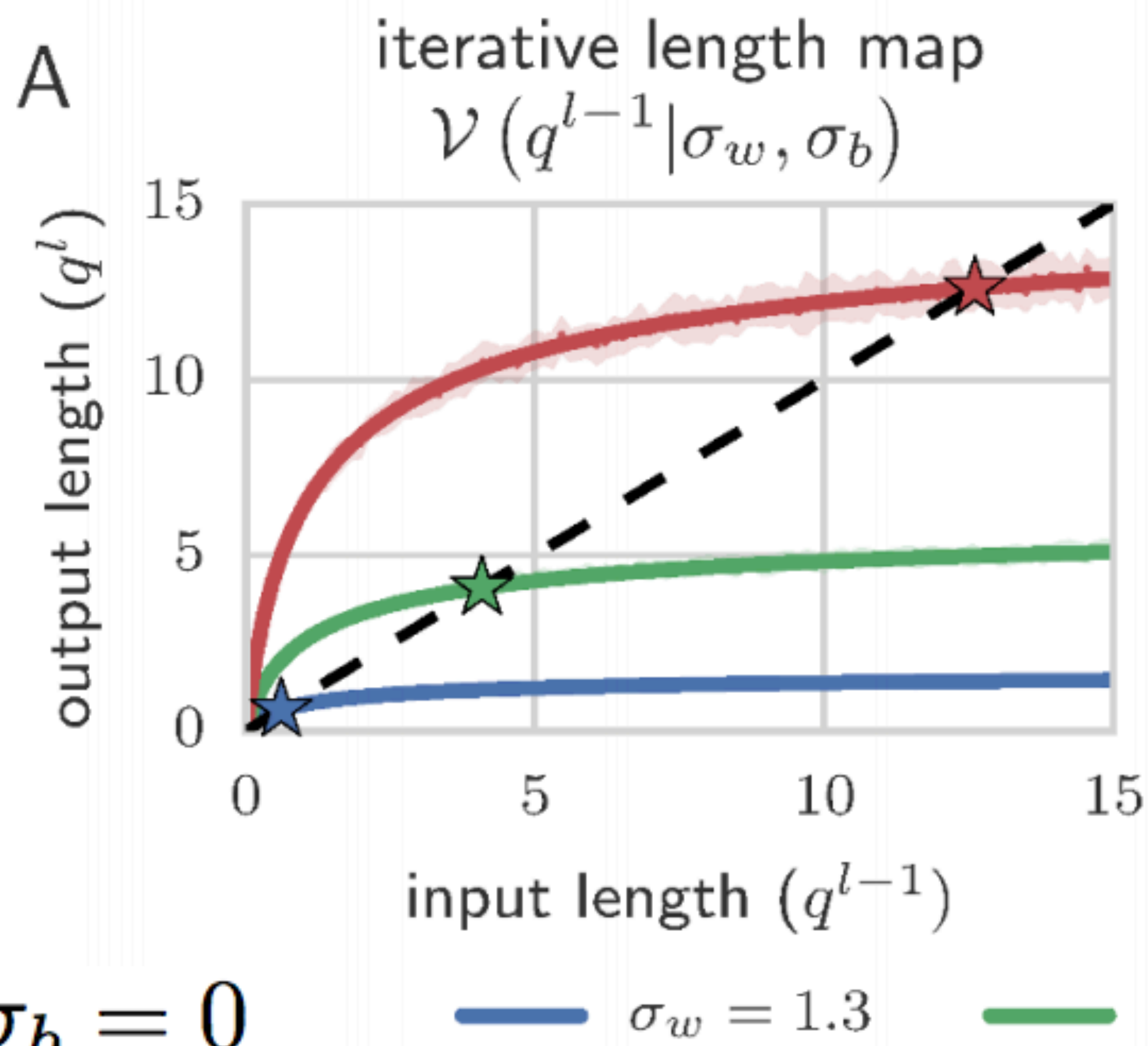
$$q^l = \mathcal{V}(q^{l-1} \,|\, \sigma_w, \sigma_b) \equiv \sigma_w^2 \int \mathcal{D}z \, \phi\left(\sqrt{q^{l-1}} z\right)^2 + \sigma_b^2, \quad \text{for} \quad l = 2, \ldots, D$$



This map can have a fixed point

Here, for: $\phi(h) = \tanh(h)$

$$q^l = \mathcal{V}(q^{l-1} \,|\, \sigma_w, \sigma_b) \equiv \sigma_w^2 \int \mathcal{D}z \, \phi\left(\sqrt{q^{l-1}} z\right)^2 + \sigma_b^2, \quad \text{for} \quad l = 2, \ldots, D$$



For $\sigma_w < 1$, all inputs shrink to zero

For $\sigma_w > 1$, the network expands small inputs and contracts large inputs

$\sigma_b = 0$

This map can have a fixed point

Here, for: $\phi(h) = \tanh(h)$

Idea 2 (Poole et al. arXiv:1606.05340): Consider two inputs $\mathbf{x}^{0,1}$ and $\mathbf{x}^{0,2}$

How will the correlation between them change while propagating through the network?

$$q^l_{ab} = \frac{1}{N_l} \sum_{i=1}^{N_l} \mathbf{h}^l_i(\mathbf{x}^{0,a})\, \mathbf{h}^l_i(\mathbf{x}^{0,b}) \qquad a, b \in \{1, 2\}$$

$$q^l_{12} = \mathcal{C}(c^{l-1}_{12},\, q^{l-1}_{11},\, q^{l-1}_{22} \,|\, \sigma_w, \sigma_b) \equiv \sigma_w^2 \int \mathcal{D}z_1\, \mathcal{D}z_2\, \phi\,(u_1)\, \phi\,(u_2) + \sigma_b^2,$$

$$u_1 = \sqrt{q^{l-1}_{11}}\, z_1, \qquad u_2 = \sqrt{q^{l-1}_{22}} \left[c^{l-1}_{12} z_1 + \sqrt{1 - (c^{l-1}_{12})^2}\, z_2 \right]$$

$$c^l_{12} = q^l_{12}(q^l_{11} q^l_{22})^{-1/2} \qquad\qquad q^l_{11} = q^l_{22} = q^*(\sigma_w, \sigma_b)$$

Idea 2 (Poole et al. arXiv:1606.05340): Consider two inputs and
How will the correlation between them change while propagating through the network?

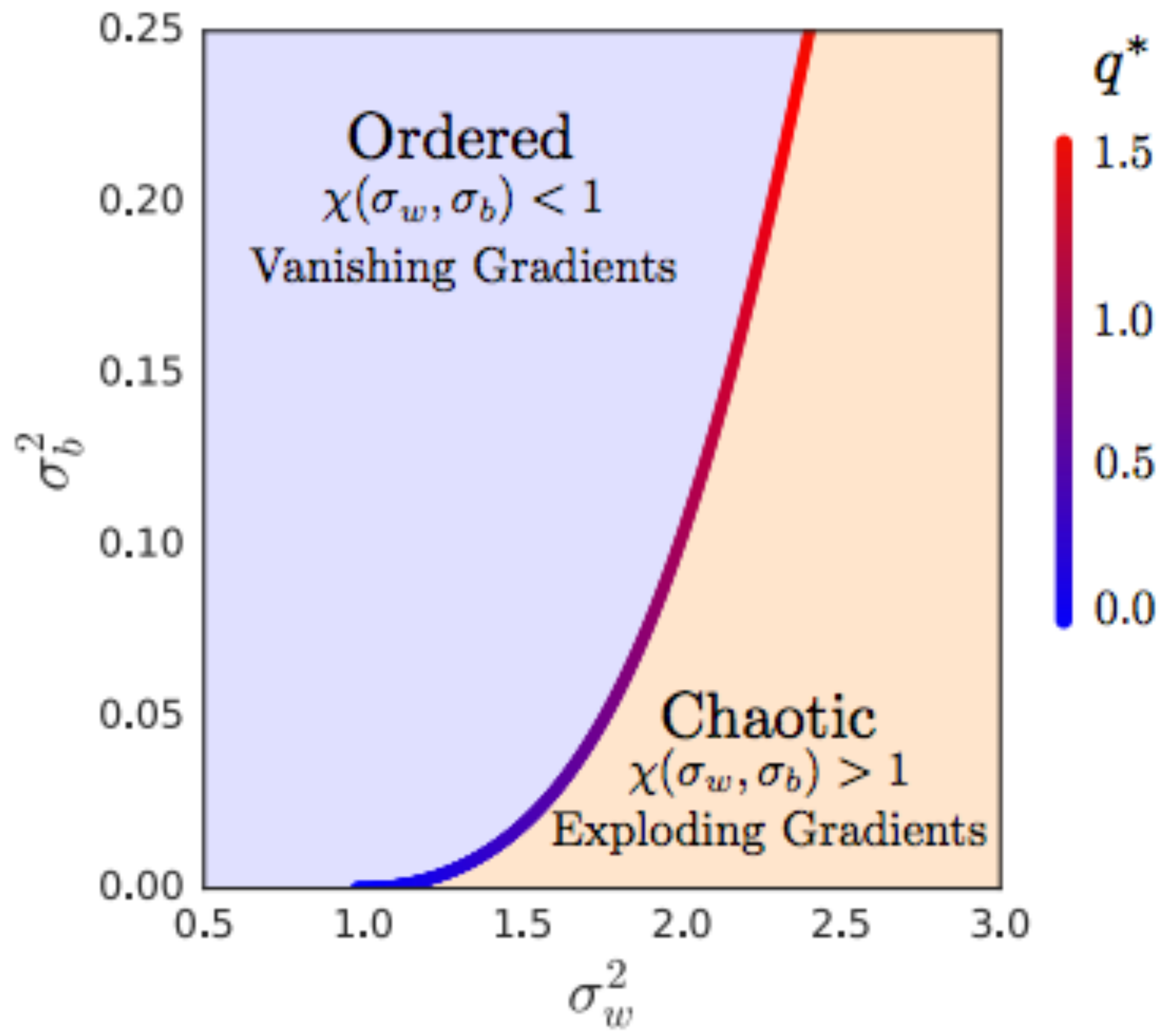
$$c_{12}^{l} = \frac{1}{q^*}\mathcal{C}(c_{12}^{l-1},\, q^*,\, q^* \,|\, \sigma_w, \sigma_b)$$

Fixed point:

$$c^* = 1$$

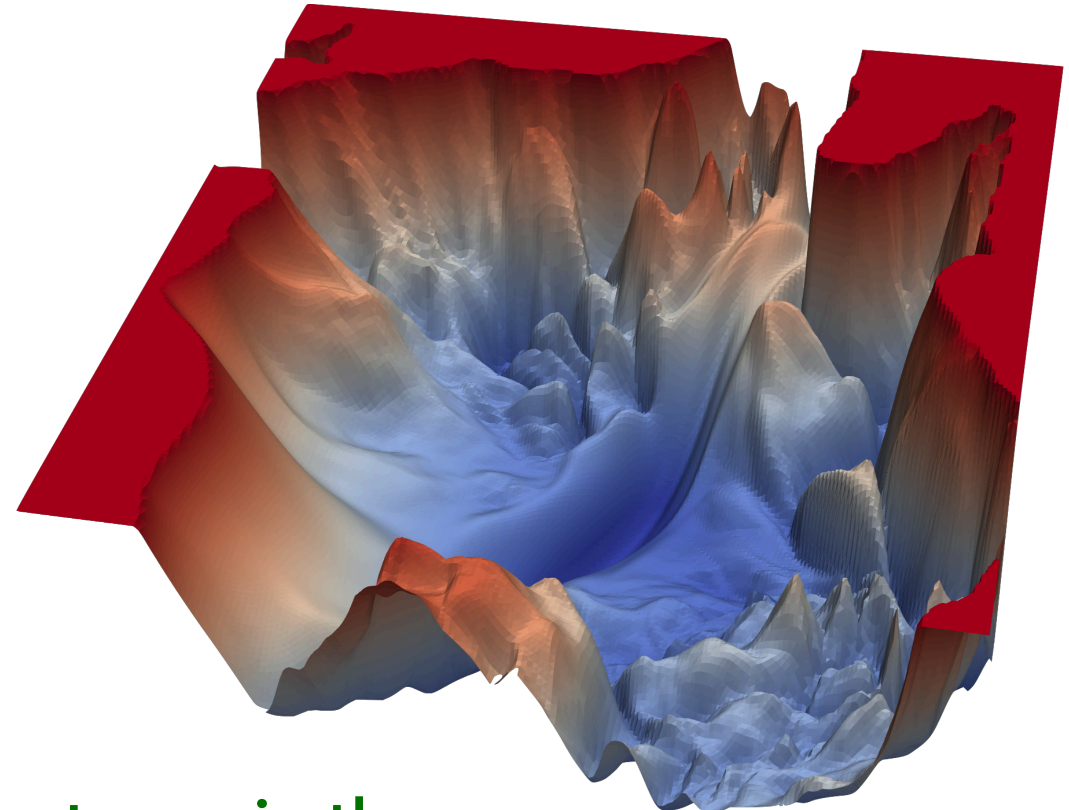
Map stability:

$$\chi_1 \equiv \left.\frac{\partial c_{12}^{l}}{\partial c_{12}^{l-1}}\right|_{c=1}$$



$$\Delta W_{ij}^{l} = -\eta \frac{\partial E(\boldsymbol{x}^L, \boldsymbol{y})}{\partial W_{ij}^{l}}$$

The learning process is based on gradually modifying the weights of the network



$$\Delta W_{ij}^{l} = -\eta \sum_{k,t} \frac{\partial x_t^l}{\partial W_{ij}^l} \frac{\partial x_k^L}{\partial x_t^l} \frac{\partial E(\boldsymbol{x}^L, \boldsymbol{y})}{\partial x_k^L}$$

All the terms in the sum of products must be bounded

$$J_{ik} = \frac{\partial x_i^L}{\partial x_k^0}$$

The input-output Jacobian is the most problematic one

It can be rewritten as:

Go to page 15

$$\boldsymbol{J} = \prod_{l=1}^{L} \boldsymbol{D}^l \boldsymbol{W}^l$$

with

$$D_{ij}^{l} = \phi'(h_i^l)\delta_{ij}$$

For a given activation function and network depth L,
how to initialize the weights?

Set the weight and bias variances, so that you're in a fixed point fro the distribution of pre-activations and on the edge of the order to chaos transition

Study Signal propagation in the network to find the statistics of

Go to page 15

$$J = \prod_{l=1}^{L} \boldsymbol{D}^l \boldsymbol{W}^l$$

with

$$D^l_{ij} = \phi'(h^l_i)\delta_{ij}$$

Use Random Matrix and Free Probability Theories to find the singular values of the Jacobian and make sure they are concentrated around 1 (Dynamical Isometry)

(Pennigton et al. arXiv:1711.04735)

Activation function: $\phi(h) = \tanh(h)$

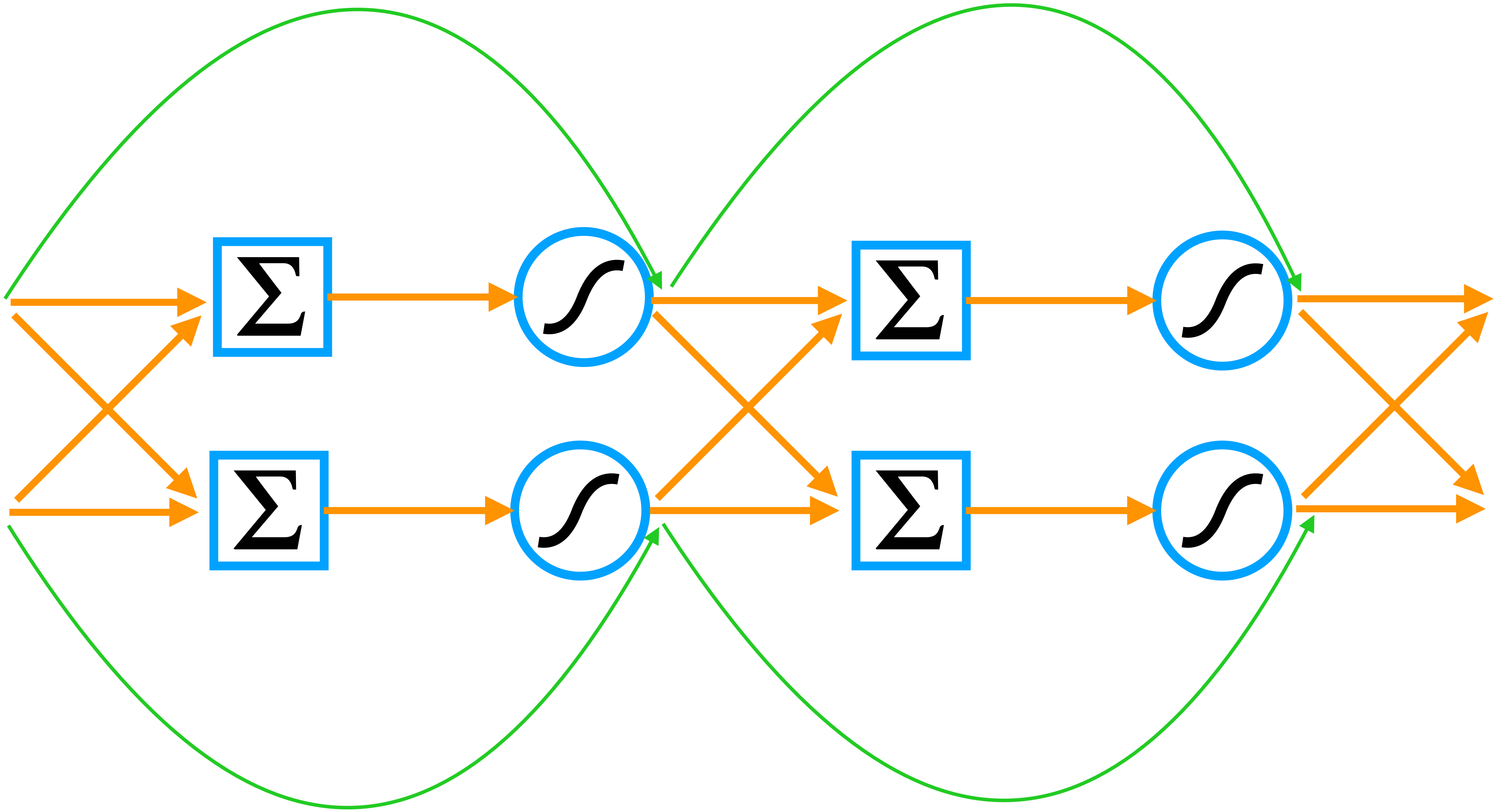
Weight matrix at initialization orthogonal: $W^T W = \mathbf{1}$

Dynamical Isometry in feed forward neural network

"Orders of magnitude" faster learning of DEEP feed forward neural networks

Not possible at all for ReLU (in feed forward networks).

Residual neural network



Signal propagation: $\mathbf{x}^{\mathbf{l}} = \phi(\mathbf{h}^{\mathbf{l}}) + \boxed{\mathbf{a}\mathbf{x}^{\mathbf{l-1}}}, \quad \mathbf{h}^{\mathbf{l}} = \mathbf{W}^{\mathbf{l}}\mathbf{x}^{\mathbf{l-1}} + \mathbf{b}^{\mathbf{l}}$

(outmatched other models in the 2015 ILSVRC and COCO competitions)

For a given activation function and network depth L,
how to initialize the weights?

Study Signal propagation in the network to find the statistics of

Are there fixed points of the maps?

Go to page 15

$$\boldsymbol{J} = \prod_{l=1}^{L} \left(\boldsymbol{D}^l \boldsymbol{W}^l + \mathbf{1}a \right), \quad \text{with} \quad D^l_{ij} = \phi'(h^l_i)\delta_{ij}$$

Use Random Matrix and Free Probability Theories to find the singular values of the Jacobian and make sure they are concentrated around 1 (Dynamical Isometry)

Signal propagation: $$\mathbf{x^l} = \phi(\mathbf{h^l}) + \mathbf{a}\mathbf{x^{l-1}}, \quad \mathbf{h^l} = \mathbf{W^l}\mathbf{x^{l-1}} + \mathbf{b^l}$$

elements of weight matrices and bias vectors - i.i.d. Gaussian with mean 0 and variances: $\sigma_w^2/(NL)$ and σ_b^2

Study: $$q^l = \frac{1}{N_l}\sum_{i=1}^{N_l}(\mathbf{h}_i^l)^2$$

The resulting mapping:

$$q^{l+1} = a^2 q^l - \left(a^2 - 1\right)\sigma_b^2 + \frac{(\sigma_W)^2}{L}\int \mathcal{D}z\phi^2\left(\sqrt{q^l}z\right) + 2\frac{(\sigma_W)^2}{L}\left[\sum_{k=1}^{l-1} a^k \int \mathcal{D}z\phi\left(\sqrt{q^{l-k}}z\right)\right]\int \mathcal{D}z\phi\left(\sqrt{q^l}z\right)$$

No fixed point. Same with the correlation map.

Same result for orthogonal weight matrices.

Go to page 15

Singular spectrum of $J = \prod_{l=1}^{L} \left(D^l W^l + \mathbf{1}a \right)$

Random Matrix Theory

$$G_H(z) = \left\langle \frac{1}{N} \mathrm{Tr} \, (z\mathbf{1} - H)^{-1} \right\rangle = \int_{-\infty}^{\infty} \frac{\rho_H(\lambda) d\lambda}{z - \lambda}$$

$$\rho_H(x) = -\frac{1}{\pi} \lim_{\epsilon \to 0} G_H(x + i\epsilon)$$

Free Probability Theory

$$G\left(R(z) + \frac{1}{z} \right) = z, \qquad R(G(z)) + \frac{1}{G(z)} = z.$$

$$R_{X+Y}(z) = R_X(z) + R_Y(z)$$

R-transform

$$S\,(zR(z)) = \frac{1}{R(z)}, \quad R(zS\,(z)) = \frac{1}{S\,(z)}.$$

$$S_{AB}(z) = S_A(z) S_B(z)$$

S-transform

Go to page 15

Singular spectrum of $J = \prod_{l=1}^{L} \left(\boldsymbol{D}^l \boldsymbol{W}^l + \mathbf{1} a \right)$

$$S_{Y_l Y_l^T}(z) = \frac{1}{a^2}\left(1 - \frac{c_2^l}{a^2 L}(1+2z) + \mathcal{O}\left(\frac{1}{L^2}\right)\right) \qquad c_2^l = \sigma_W^2 \left\langle (\phi'(h))^2 \right\rangle_l = \sigma_W^2 \int \mathcal{D}z \phi'^2\left(\sqrt{q^l} z\right)$$

define the effective cumulant: $c = \frac{1}{L}\sum_{l=1}^{L} c_2^l$

The large network depth limit (recall the scaling of the variance: $\sigma_w^2/(NL)$)

$$\ln S_{JJ^T}(z) = -2L\ln a + \sum_{l=1}^{L} \ln\left(1 - \frac{c_2^l}{a^2 L}(1+2z)\right) \approx -2L\ln a - \frac{1+2z}{a^2 L}\sum_{l=1}^{L} c_2^l =: -2L\ln a - \frac{(1+2z)}{a^2} c,$$

$$S_{JJ^T}(z) = \frac{1}{a^{2L}} e^{-\frac{c}{a^2}(1+2z)}.$$

Universal formula for any activation function!

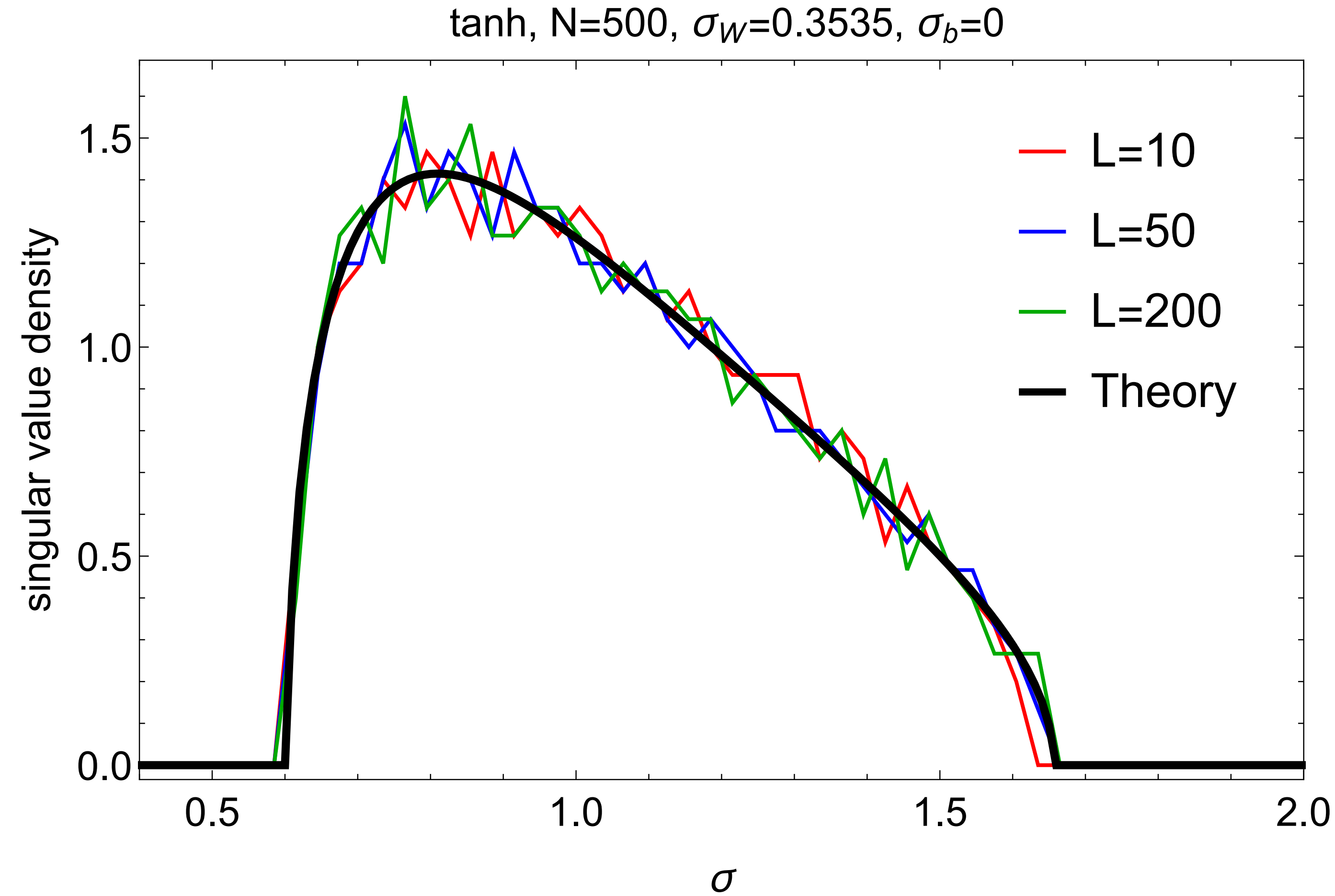
$$a^{2L} G(z) = (zG(z) - 1) e^{\frac{c}{a^2}(1 - 2zG(z))}$$

(solution in terms of the Lambert function)

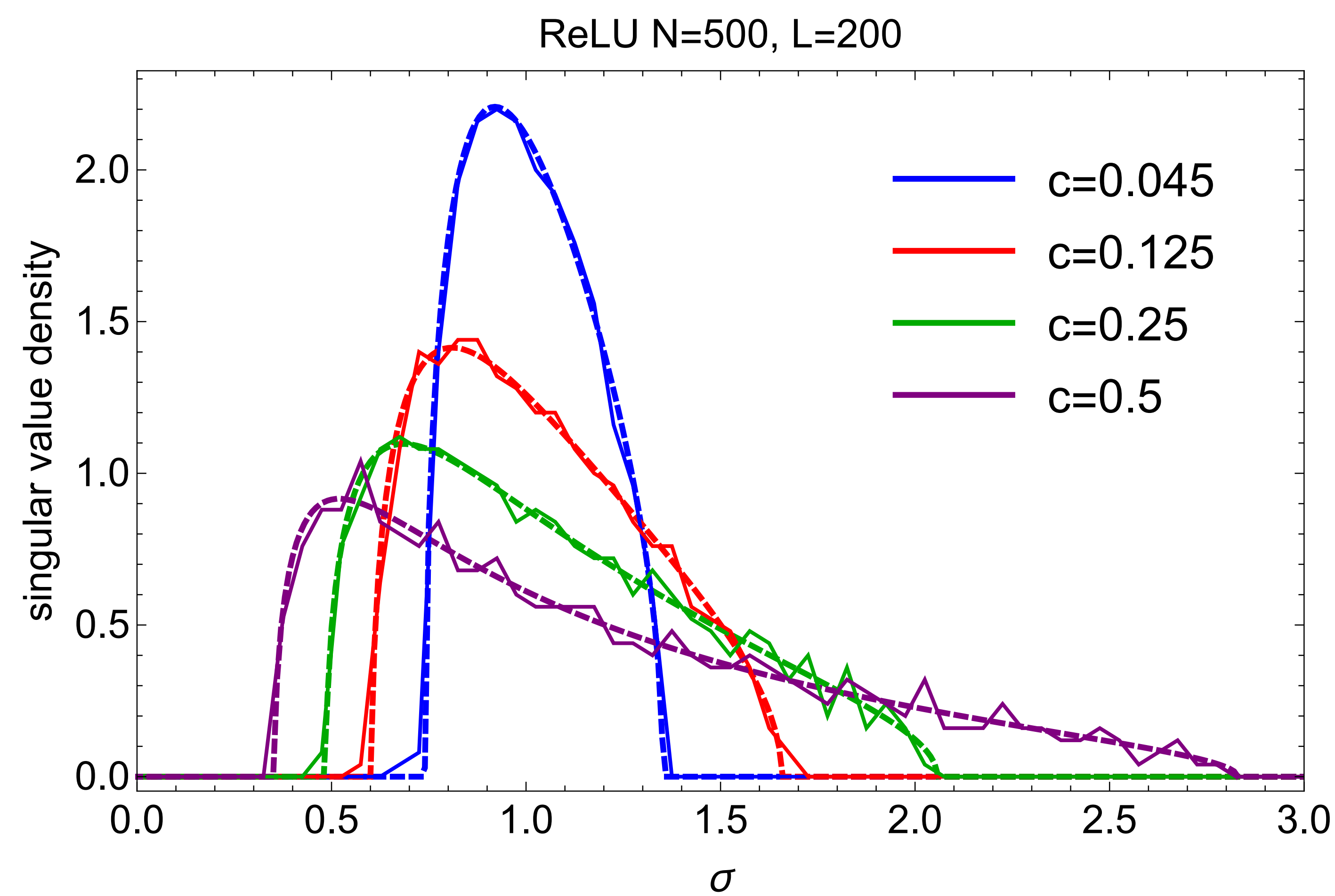
$$c_2^l = \left\langle \frac{1}{N} \mathrm{Tr} \boldsymbol{W}^l \boldsymbol{D}^l \boldsymbol{D}^l (\boldsymbol{W}^l)^T \right\rangle = \frac{\sigma_w^2}{N} \sum_i^N \left(\phi'(h_i^l)\right)^2 = \sigma_W^2 \int \mathcal{D}z \phi'^2\left(\sqrt{q^l} z\right)$$

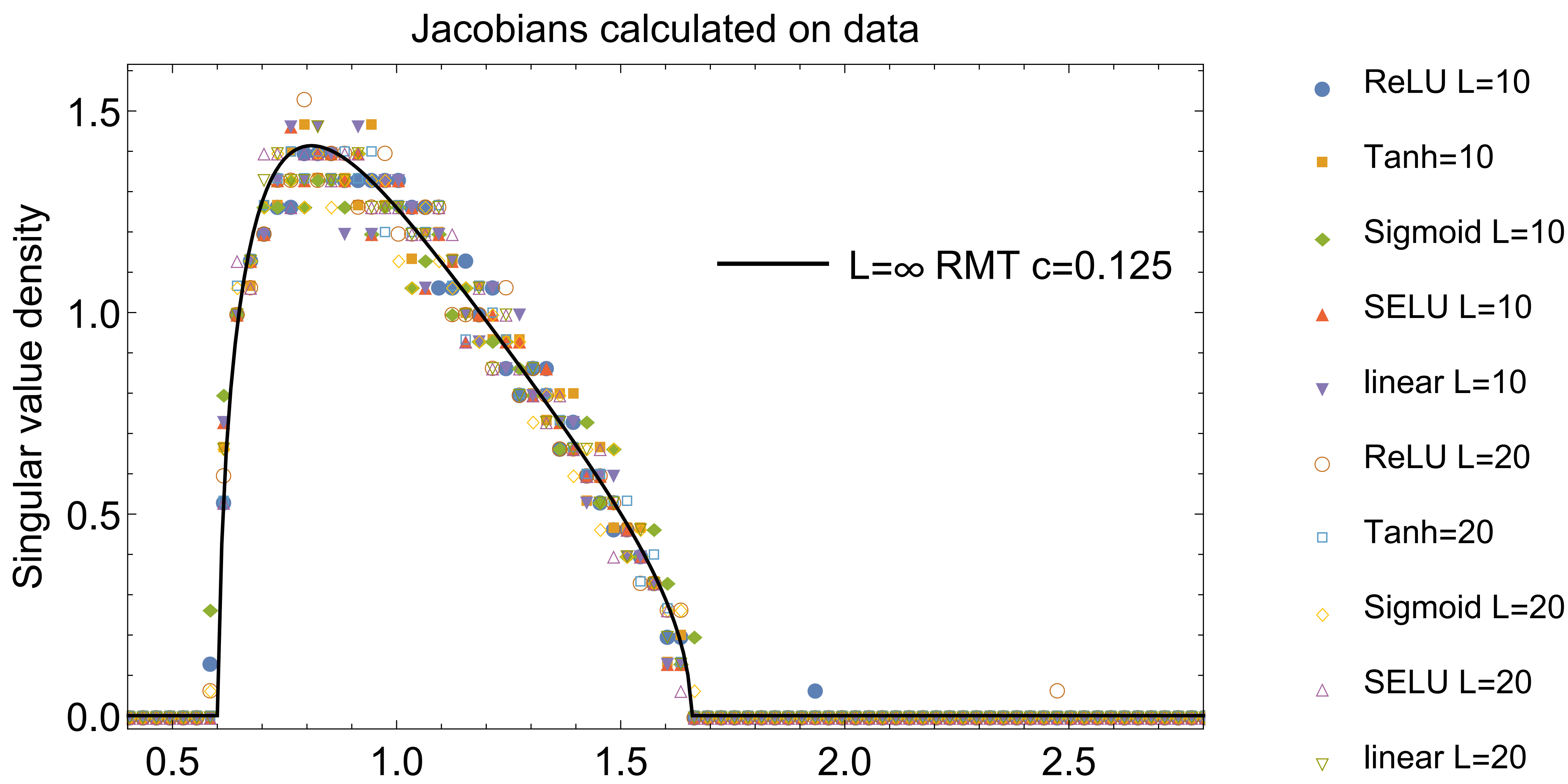
With a proper scaling of the variances of the weights, the result is a universal formula for the probability density of the singular values, depending on a single parameter c.

With a proper scaling of the variances of the weights, the result is a universal formula for the probability density of the singular values, depending on a single parameter c.

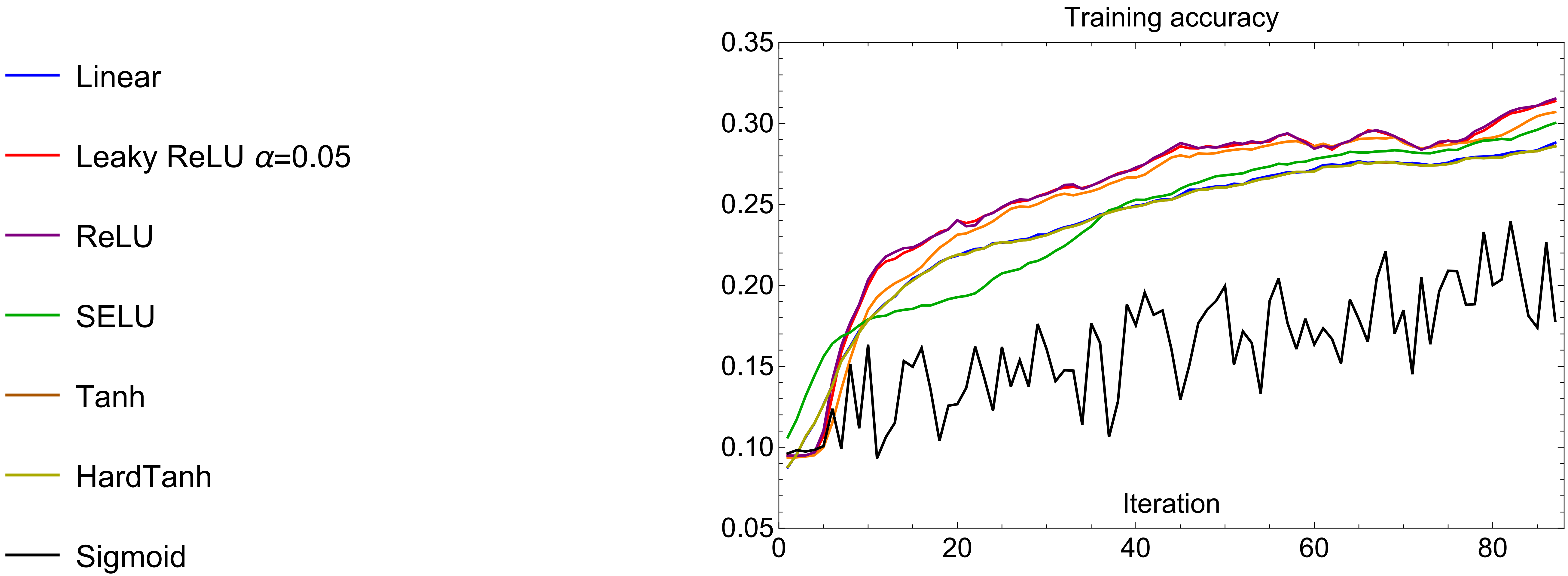


Corroborated with numerical experiments with random matrices.

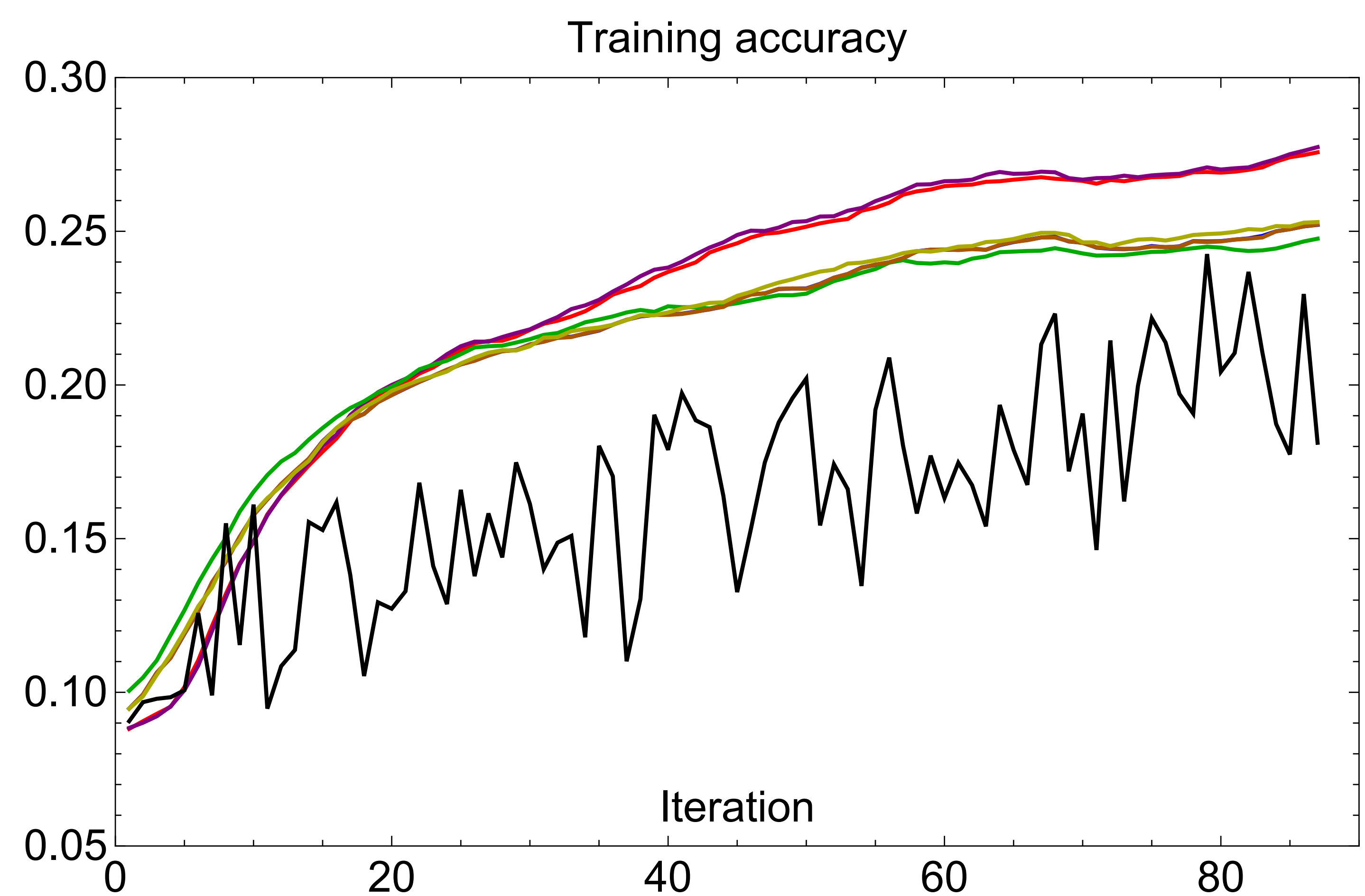




Corroborated with numerical experiments with neural networks.



These results allow us to eliminate the singular spectrum of the Jacobian treated as a confounding factor in experiments with the learning process of simple residual neural networks for different activation functions enabling meaningful comparisons.



These results published on arXiv:1809.08848

Thank you for your attention.

JAGIELLONIAN UNIVERSITY
IN KRAKÓW