Theories of deep learning: generalization, expressivity, and training

Surya Ganguli

Dept. of Applied Physics,
Neurobiology,
and Electrical Engineering

Stanford University

Funding:

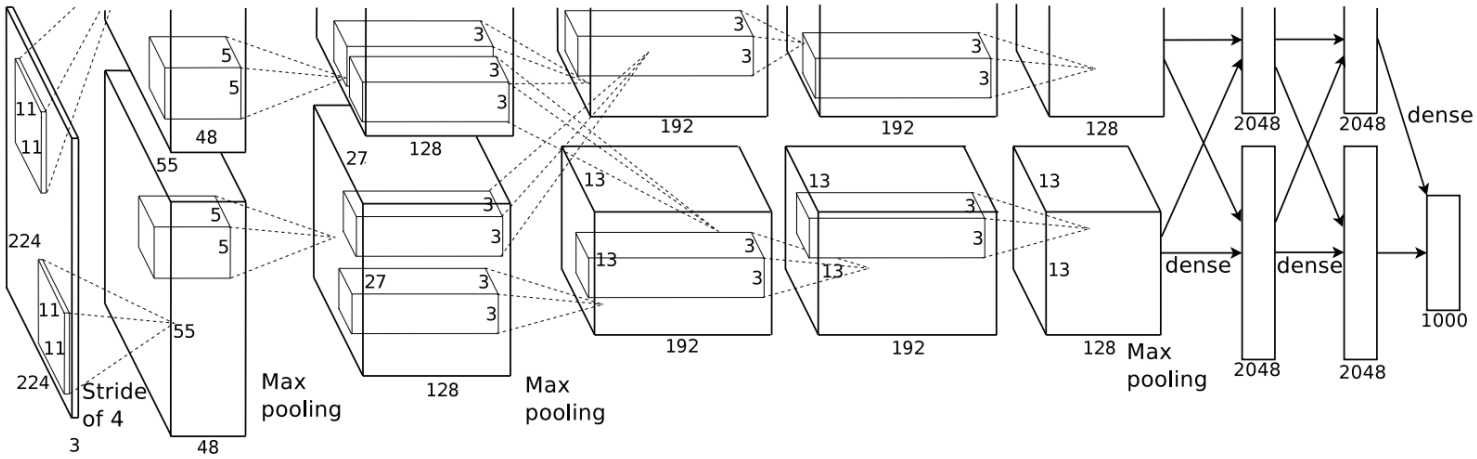
Bio-X Neuroventures
Burroughs Wellcome
Genentech Foundation
James S. McDonnell Foundation
McKnight Foundation
National Science Foundation

NIH
Office of Naval Research
Simons Foundation
Sloan Foundation
Swartz Foundation
Stanford Terman Award

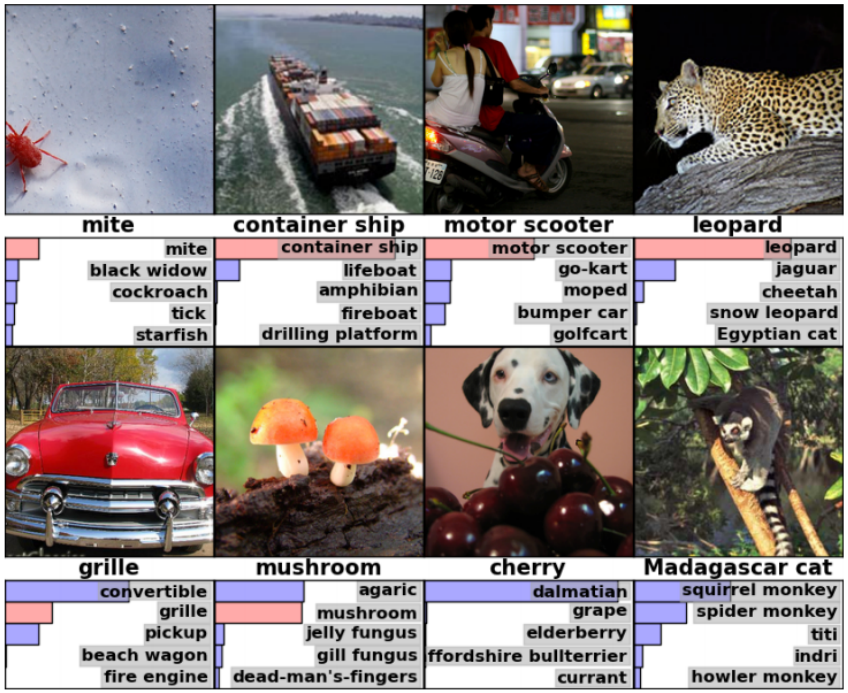
http://ganguli-gang.stanford.edu

Twitter: @SuryaGanguli

An interesting artificial neural circuit for image classification



Alex Krizhevsky
Ilya Sutskever
Geoffrey E. Hinton
NIPS 2012



References: http://ganguli-gang.stanford.edu

- M. Advani and S. Ganguli, An equivalence between high dimensional Bayes optimal inference and M-estimation, NIPS 2016.
- M. Advani and S. Ganguli, Statistical mechanics of optimal convex inference in high dimensions, Physical Review X, 6, 031034, 2016.
- A. Saxe, J. McClelland, S. Ganguli, Learning hierarchical category structure in deep neural networks, Proc. of the 35th Cognitive Science Society, pp. 1271-1276, 2013.
- A. Saxe, J. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep neural networks, ICLR 2014.
- Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, NIPS 2014.
- B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, NIPS 2016.
- S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, Deep information propagation, ICLR 2017.
- S. Lahiri, J. Sohl-Dickstein and S. Ganguli, A universal tradeoff between energy speed and accuracy in physical communication, arxiv 1603.07758
- A memory frontier for complex synapses, S. Lahiri and S. Ganguli, NIPS 2013.
- Continual learning through synaptic intelligence, F. Zenke, B. Poole, S. Ganguli, ICML 2017.
- Modelling arbitrary probability distributions using non-equilibrium thermodynamics, J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, ICML 2015.
- Deep Knowledge Tracing, C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas, J. Sohl-Dickstein, NIPS 2015.
- Deep learning models of the retinal response to natural scenes, L. McIntosh, N. Maheswaranathan, S. Ganguli, S. Baccus, NIPS 2016.
- Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice, J. Pennington, S. Schloenholz, and S. Ganguli, NIPS 2017.
- Variational walkback: learning a transition operator as a recurrent stochastic neural net, A. Goyal, N.R. Ke, S. Ganguli, Y. Bengio, NIPS 2017.
- The emergence of spectral universality in deep networks, J. Pennington, S. Schloenholz, and S. Ganguli, AISTATS 2018.

Tools: Non-equilibrium statistical mechanics
Dynamical mean field theory
Statistical mechanics of random landscapes
Riemannian geometry
Random matrix theory
Free probability theory

At a coarse grained level: 3 puzzles of deep learning

Generalization: How can neural networks predict the response to new examples?

A. Saxe, J. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep neural networks, ICLR 2014.

A. Lampinen, J. McCelland, S. Ganguli, An analytic theory of generalization dynamics and transfer learning in deep linear networks, work in progress.

Expressivity: Why deep? What can a deep neural network "say" that a shallow network cannot?

B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, NIPS 2016.

Trainability: How can we optimize non-convex loss functions to achieve small training error?

Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice, J. Pennington, S. Schloenholz, and S. Ganguli, NIPS 2017.

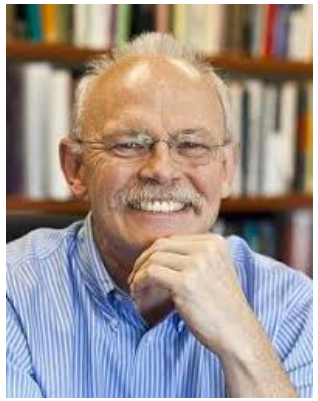
The emergence of spectral universality in deep networks, J. Pennington, S. Schloenholz, and S. Ganguli, AISTATS 2018.

Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, NIPS 2014.

Learning dynamics of training and testing error

Andrew Saxe
Harvard

Andrew Lampinen
Stanford

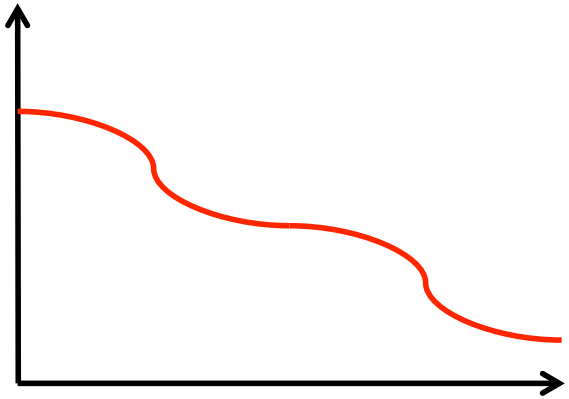


Jay McClelland
Stanford

Learning dynamics of training and testing error

The dynamics of learning in deep nonlinear networks is complex:

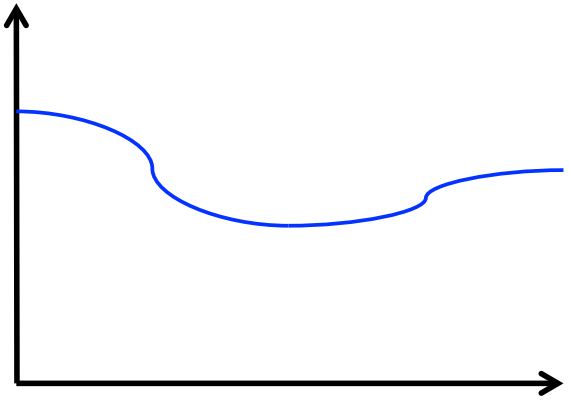
Training error



Training time

Plateaus with sudden
sudden transitions to lower error

Test error

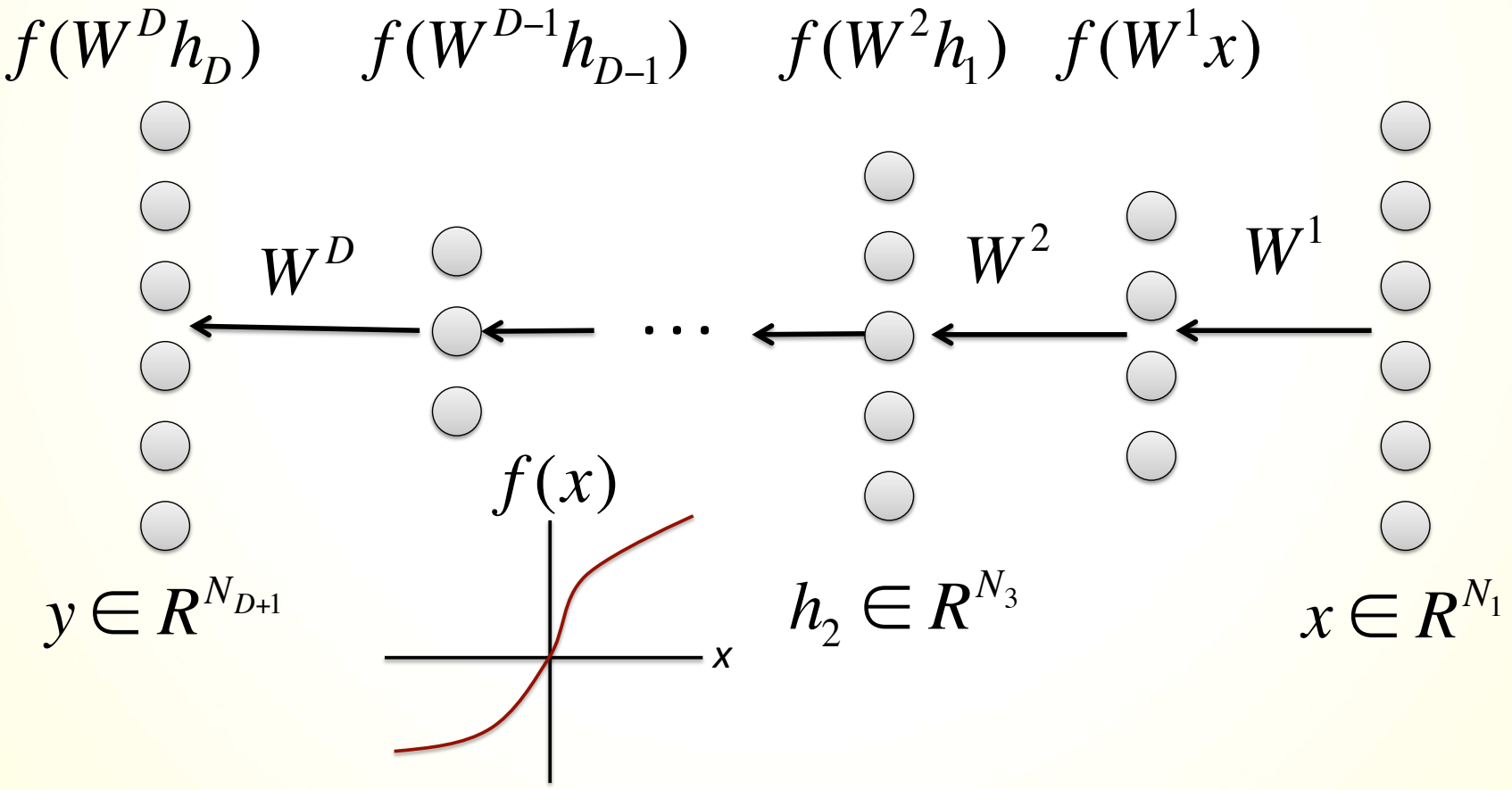


Training time

Non-monotonicity:
Overfitting to training examples;
Bad predictions on new examples

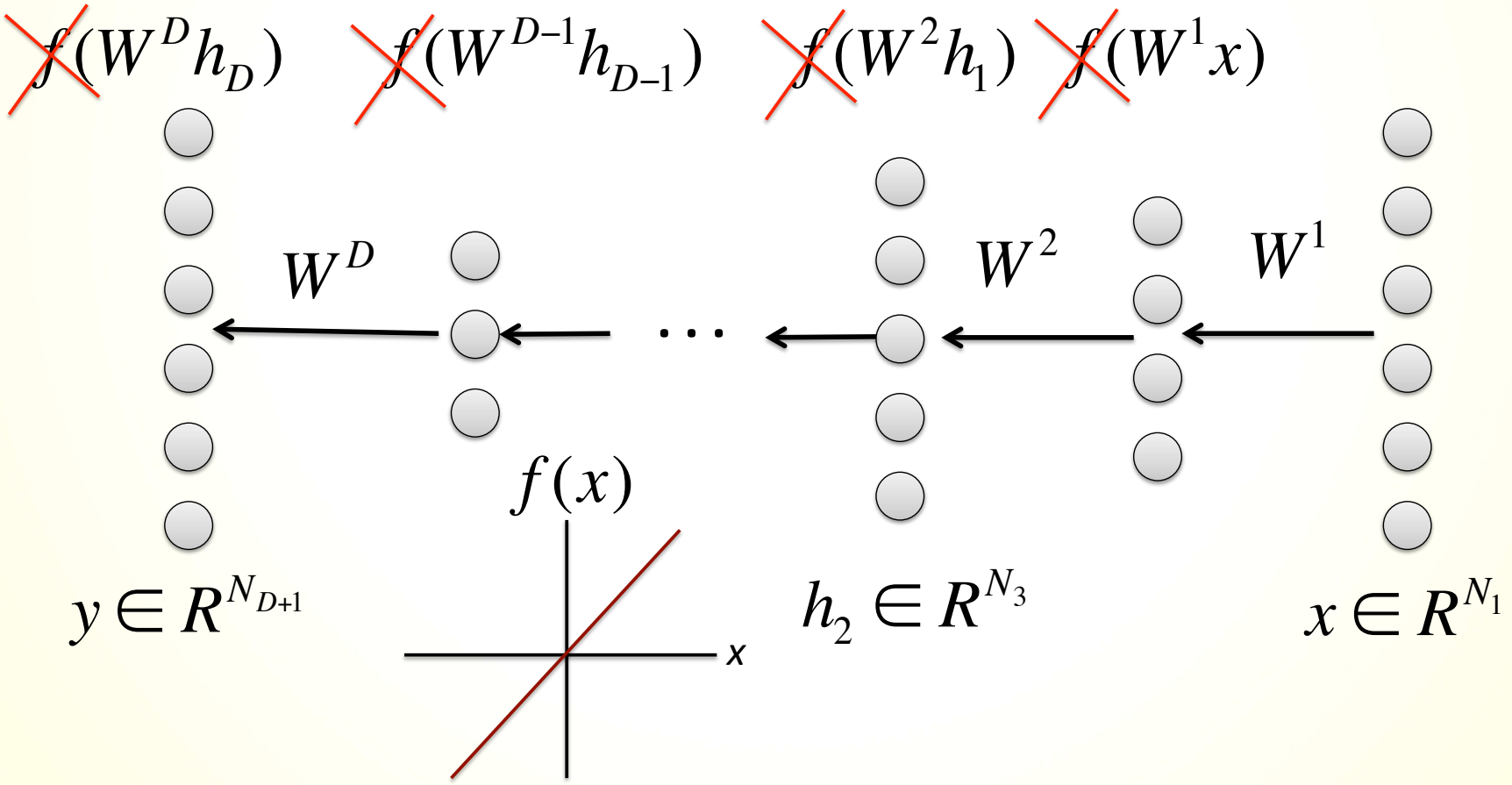
Deep network

- Little hope for a complete theory with arbitrary nonlinearities



Deep *linear* network

- Use a deep *linear* network as a starting point



Deep *linear* network

- Input-output map: Always linear

$$y = \left(\prod_{i=1}^{D} W^i \right) x \equiv W^{tot} x$$

- Gradient descent dynamics: Nonlinear; coupled; nonconvex

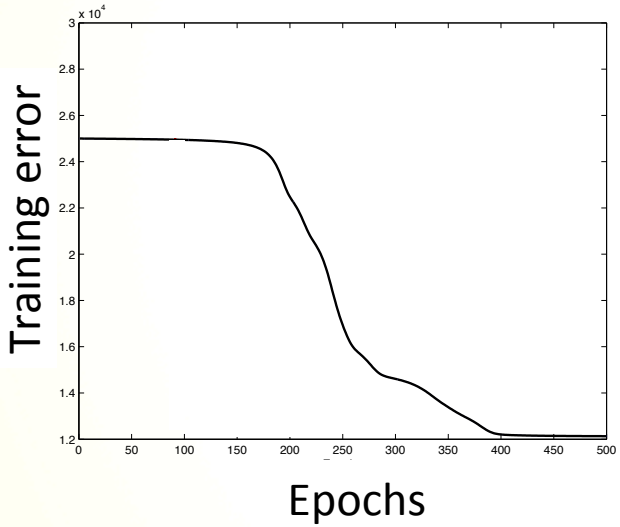
$$\Delta W^l = \lambda \sum_{\mu=1}^{P} \left(\prod_{i=l+1}^{D} W^i \right)^T \left[y^\mu x^{\mu T} - \left(\prod_{i=1}^{D} W^i \right) x^\mu x^{\mu T} \right] \left(\prod_{i=1}^{l-1} W^i \right)^T$$

$$l = 1, \cdots, D$$

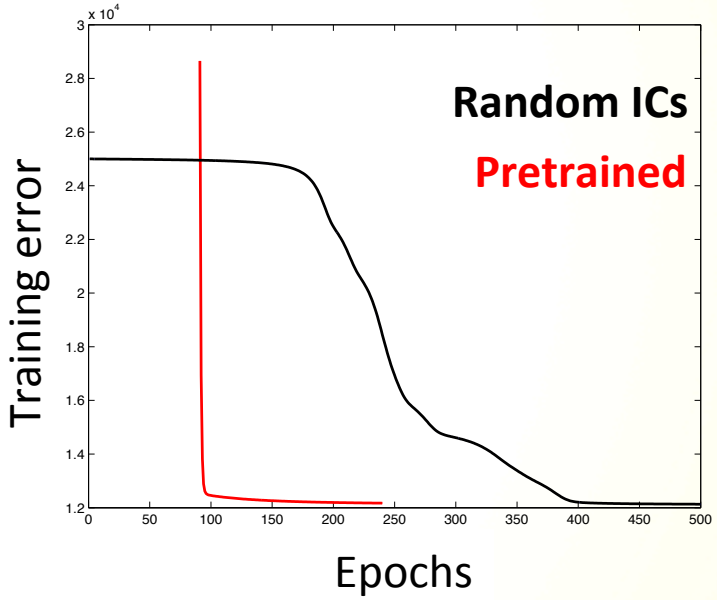
- Useful for studying *learning dynamics*, not representation power.

Nontrivial learning dynamics

Plateaus and sudden transitions

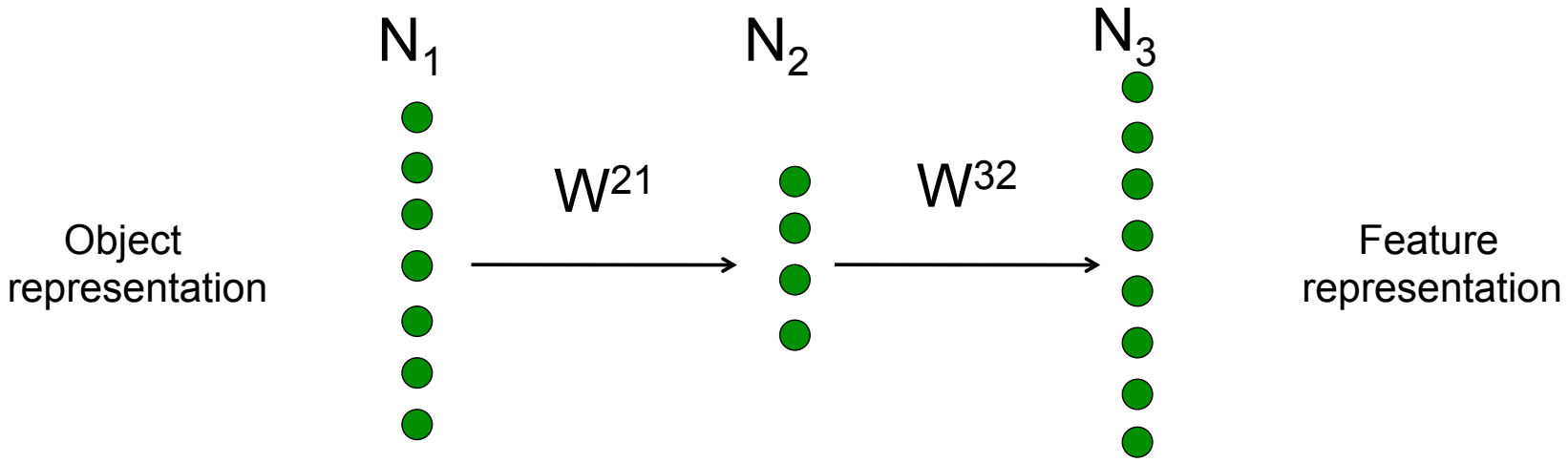


Faster convergence from pretrained initial conditions



- Build intuitions for nonlinear case by analyzing linear case

Nonlinear learning dynamics in a 3 layer linear net



$\mathbf{W}^{21}$ is an $N_2 \times N_1$ matrix of synaptic connections from layer 1 to 2
$\mathbf{W}^{32}$ is an $N_3 \times N_2$ matrix of connections from layer 2 to 3.

The input-output map of the network is $\mathbf{y} = \mathbf{W}^{32}\mathbf{W}^{21}\mathbf{x}$.

P training examples $\{\mathbf{x}^\mu, \mathbf{y}^\mu\}, \mu = 1, \ldots, P$,

Learning = gradient descent on squared error: $||\mathbf{y}^\mu - \mathbf{W}^{32}\mathbf{W}^{21}\mathbf{x}^\mu||^2$

$$\Delta \mathbf{W}^{21} = \lambda \mathbf{W}^{32^T}\left(\mathbf{y}^u\mathbf{x}^{\mu T} - \mathbf{W}^{32}\mathbf{W}^{21}\,\mathbf{x}^\mu\mathbf{x}^{\mu T}\right)$$
$$\Delta \mathbf{W}^{32} = \lambda \left(\mathbf{y}^u\mathbf{x}^{\mu T} - \mathbf{W}^{32}\mathbf{W}^{21}\,\mathbf{x}^\mu\mathbf{x}^{\mu T}\right)\mathbf{W}^{21^T}$$

Averaging over the input statistics

$$\tau \frac{d}{dt} \mathbf{W}^{21} = \mathbf{W}^{32T} \left(\mathbf{\Sigma}^{31} - \mathbf{W}^{32} \mathbf{W}^{21} \mathbf{\Sigma}^{11} \right)$$
$$\tau \frac{d}{dt} \mathbf{W}^{32} = \left(\mathbf{\Sigma}^{31} - \mathbf{W}^{32} \mathbf{W}^{21} \mathbf{\Sigma}^{11} \right) \mathbf{W}^{21T},$$

where

$$\mathbf{\Sigma}^{11} \equiv \sum_{\mu=1}^{P} \mathbf{x}^{\mu} \mathbf{x}^{\mu T}$$

is an $N_1 \times N_1$ input correlation matrix,

$$\mathbf{\Sigma}^{31} \equiv \sum_{\mu=1}^{P} \mathbf{y}^{\mu} \mathbf{x}^{\mu T},$$

is an $N_3 \times N_1$ input-output correlation matrix, and

$$\tau \equiv \frac{P}{\lambda}.$$

Input statistics guide change of synaptic coordinates

Change of coordinates on synaptic weight space through SVD of the input-output correlation matrix:

$$\mathbf{\Sigma}^{31} = \mathbf{U}^{33}\mathbf{S}^{31}\mathbf{V}^{11^T} = \sum_{\alpha=1}^{N_1} s_\alpha \mathbf{u}_\alpha \mathbf{v}_\alpha^T,$$

$\mathbf{V}^{11}$ is an $N_1 \times N_1$ orthogonal matrix whose columns $\mathbf{v}_\alpha$ reflect independent modes of variation in the input

$\mathbf{U}^{33}$ is an $N_3 \times N_3$ orthogonal matrix whose columns $\mathbf{u}_\alpha$ reflect independent modes of variation in the output

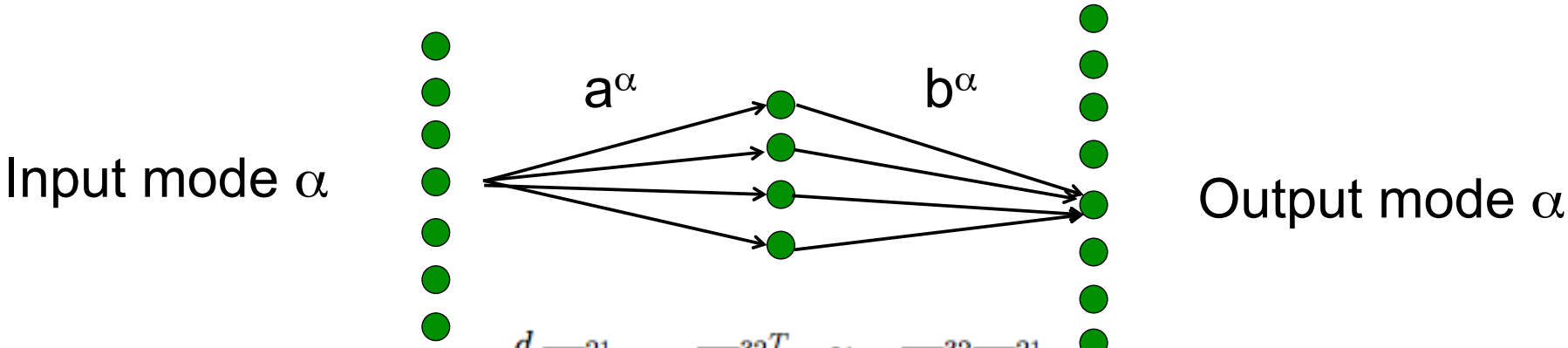
$\mathbf{S}^{31}$ is an $N_3 \times N_1$ matrix whose only nonzero elements are on the diagonal; these elements are the singular values s_α, $\alpha = 1, \ldots, N_1$ ordered so that $s_1 \geq s_2, \ldots, \geq s_{N_1}$.

We now perform the following change of variables on synaptic weight space,

$$\begin{aligned} \mathbf{W}^{21} &= \overline{\mathbf{W}}^{21}\mathbf{V}^{11^T} \\ \mathbf{W}^{32} &= \mathbf{U}^{11}\overline{\mathbf{W}}^{32}. \end{aligned}$$

$$\begin{aligned} \tau \frac{d}{dt}\overline{\mathbf{W}}^{21} &= \overline{\mathbf{W}}^{32^T}\left(\mathbf{S}^{31} - \overline{\mathbf{W}}^{32}\overline{\mathbf{W}}^{21}\right) \\ \tau \frac{d}{dt}\overline{\mathbf{W}}^{32} &= \left(\mathbf{S}^{31} - \overline{\mathbf{W}}^{32}\overline{\mathbf{W}}^{21}\right)\overline{\mathbf{W}}^{21^T}, . \end{aligned}$$

Dynamics of synaptic modes



$$\tau \frac{d}{dt} \overline{\overline{\mathbf{W}}}^{21} = \overline{\overline{\mathbf{W}}}^{32T} \left(\mathbf{S}^{31} - \overline{\overline{\mathbf{W}}}^{32} \overline{\overline{\mathbf{W}}}^{21} \right)$$

$$\tau \frac{d}{dt} \overline{\overline{\mathbf{W}}}^{32} = \left(\mathbf{S}^{31} - \overline{\overline{\mathbf{W}}}^{32} \overline{\overline{\mathbf{W}}}^{21} \right) \overline{\overline{\mathbf{W}}}^{21T},.$$

Intuition:

$\overline{\overline{\mathbf{W}}}^{21}{}_{i\alpha}$ connects $\mathbf{v}_\alpha$ to hidden node i. Let $\mathbf{a}^\alpha$ be the α'th column of $\overline{\overline{\mathbf{W}}}^{21}$
$\overline{\overline{\mathbf{W}}}^{32}{}_{\alpha i}$ connects i to output mode $\mathbf{u}_\alpha$. Let $\mathbf{b}^{\alpha T}$ be the α'th row of $\overline{\overline{\mathbf{W}}}^{32}$.

$$\tau \frac{d}{dt} \mathbf{a}^\alpha = s_\alpha \mathbf{b}^\alpha - (\mathbf{a}^\alpha \cdot \mathbf{b}^\alpha) \mathbf{b}^\alpha - \sum_{\gamma \neq \alpha}^{N_2} \mathbf{b}^\gamma (\mathbf{a}^\alpha \cdot \mathbf{b}^\gamma)$$

$$\tau \frac{d}{dt} \mathbf{b}^\alpha = s_\alpha \mathbf{a}^\alpha - (\mathbf{a}^\alpha \cdot \mathbf{b}^\alpha) \mathbf{a}^\alpha - \sum_{\gamma \neq \alpha}^{N_1} \mathbf{a}^\gamma (\mathbf{b}^\alpha \cdot \mathbf{a}^\gamma).$$

Cooperative growth ↗ Stabilization ↑ Inter-mode competition ↖

Fixed points

- As $t \longrightarrow \infty$, weights approach

$$W^{32}(t)W^{21}(t) \quad \longrightarrow \quad \Sigma^{31} = U^{33}S^{31}V^{11\,T} = \sum_{\alpha=1}^{N_1} s_\alpha u^\alpha v^{\alpha T}$$

- (Baldi & Hornik, 1989; Sanger, 1989)

- Simple end point
- What *dynamics* occur along the way?

Analytic learning trajectory

SVD of input-output correlations:

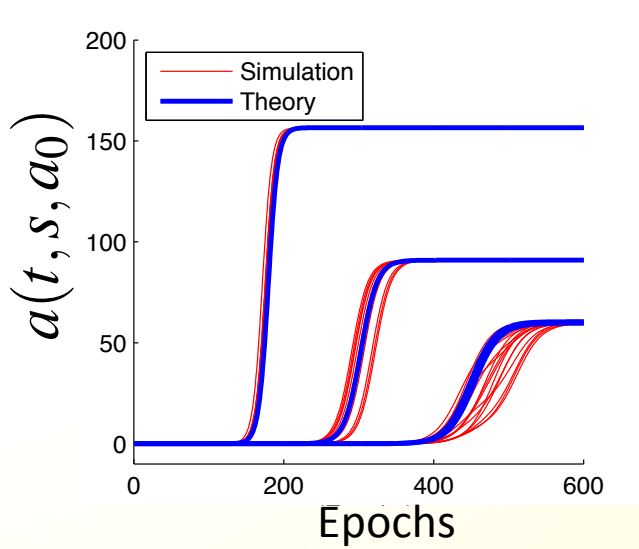
$$\Sigma^{31} = U^{33} S^{31} V^{11\,T} = \sum_{\alpha=1}^{N_1} s_\alpha u^\alpha v^{\alpha T}$$

τ	1/Learning rate
s	Singular value
a_0	Initial mode strength

Network input-output map:

$$W^{32}(t) W^{21}(t) = \sum_{\alpha=1}^{N_2} a(t, s_\alpha, a_\alpha^0)\, u^\alpha v^{\alpha T} \quad \text{where} \quad a(t,s,a_0) = \frac{s e^{2st/\tau}}{e^{2st/\tau} - 1 + s/a_0}$$

- Starting from decoupled initial conditions.
- Each 'connectivity mode' evolves independently
- Singular value s learned at time O(1/s)



Saxe, McCelland, Ganguli, ICLR, 2014

Deeper networks

- Can generalize to arbitrary depth network

- Each effective singular value *a* evolves independently

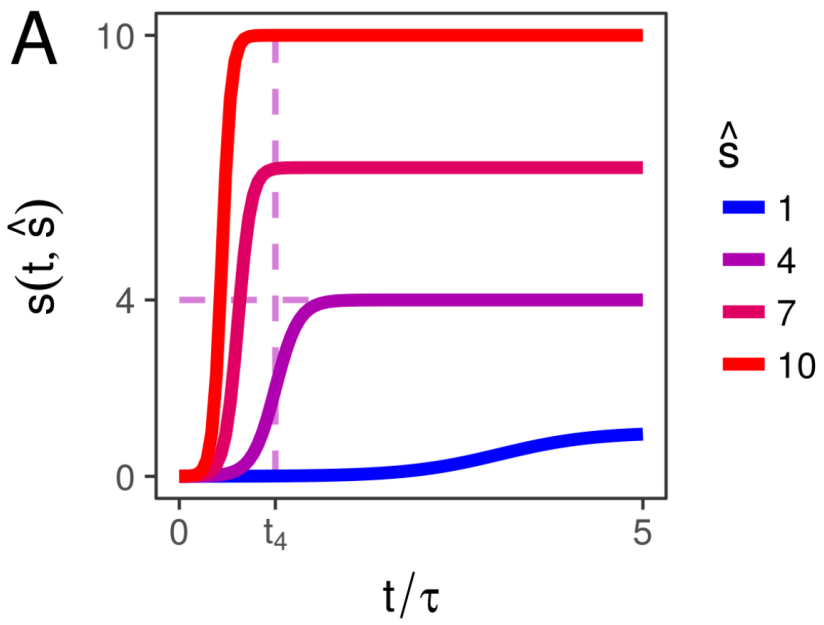
$$\tau \frac{d}{dt} a = (N_l - 1) a^{2-2/(N_l-1)} (s - a)$$

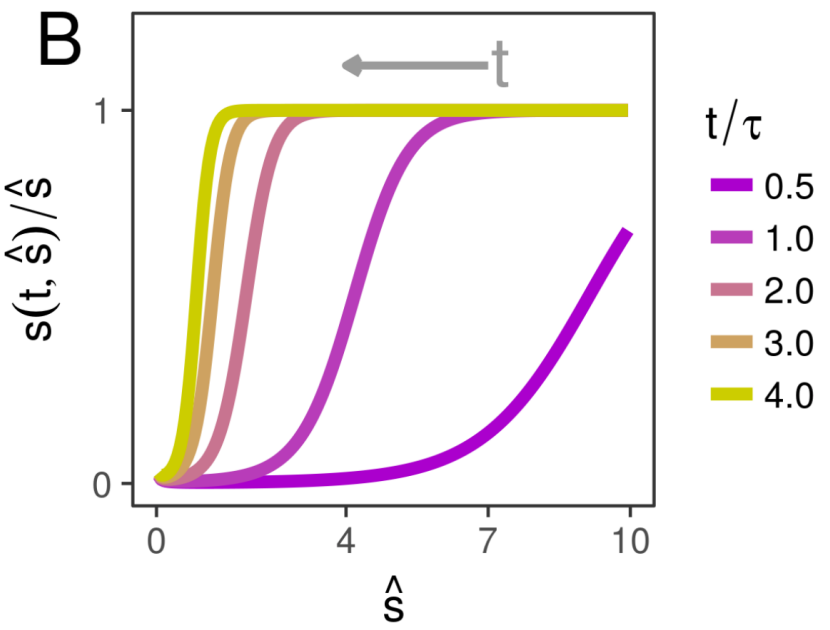
τ	1/Learning rate
s	Singular value
N_l	# layers

- In deep networks, combined gradient is $O(N_l/\tau)$

$$a = \prod_{i=1}^{N_l - 1} W_i$$

Learning as a singular mode detection wave





$$s(t,\hat{s}) = \frac{\hat{s}e^{2\hat{s}t/\tau}}{e^{2\hat{s}t/\tau} - 1 + \hat{s}/\epsilon},$$

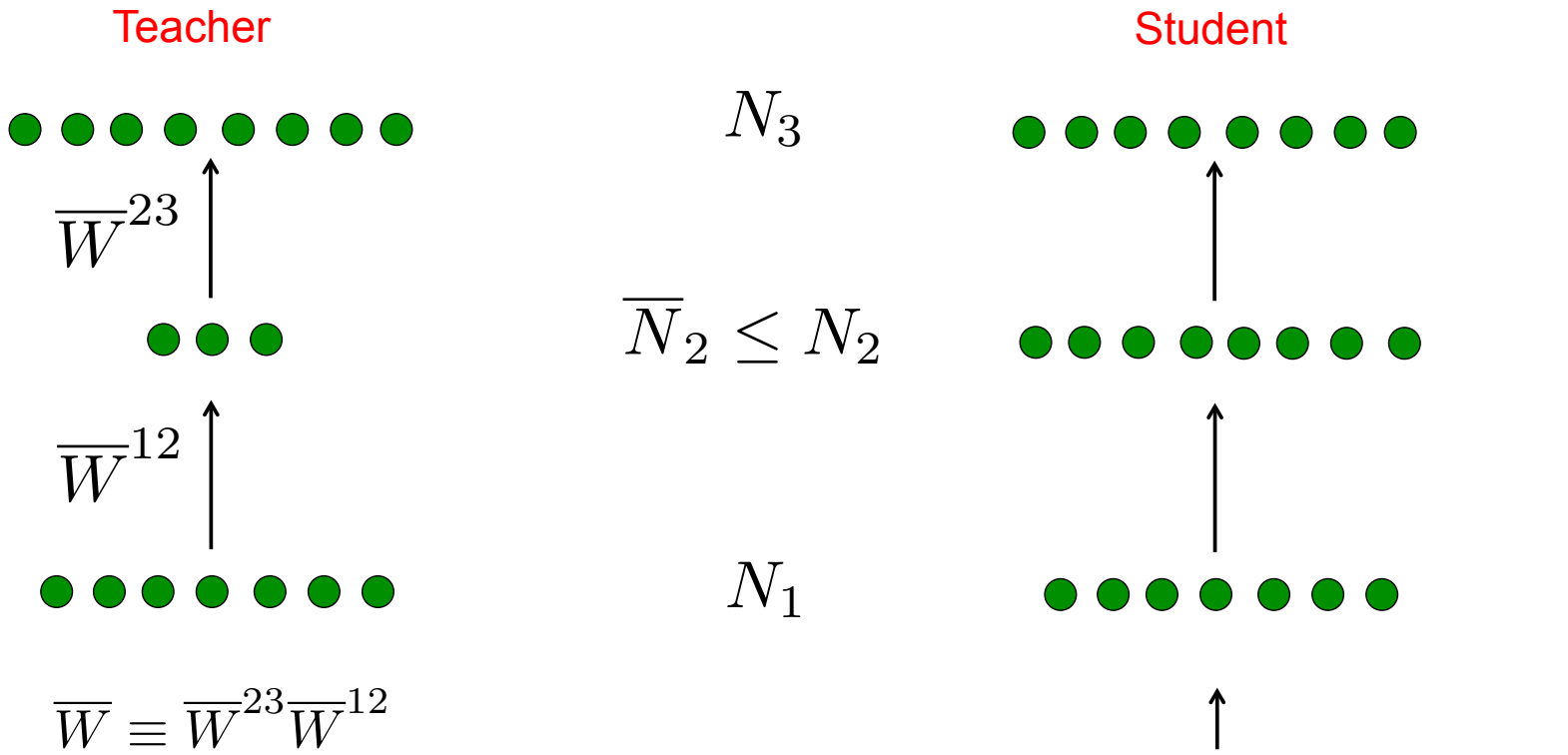
$$t(s,\hat{s}) = \frac{\tau}{2\hat{s}} \ln \frac{\hat{s}/\epsilon - 1}{\hat{s}/s - 1}$$

At time t, data singular modes with:

$\hat{s} > t/\tau$ are learned.

$\hat{s} < t/\tau$ are not yet learned.

Generalization in student teacher scenario



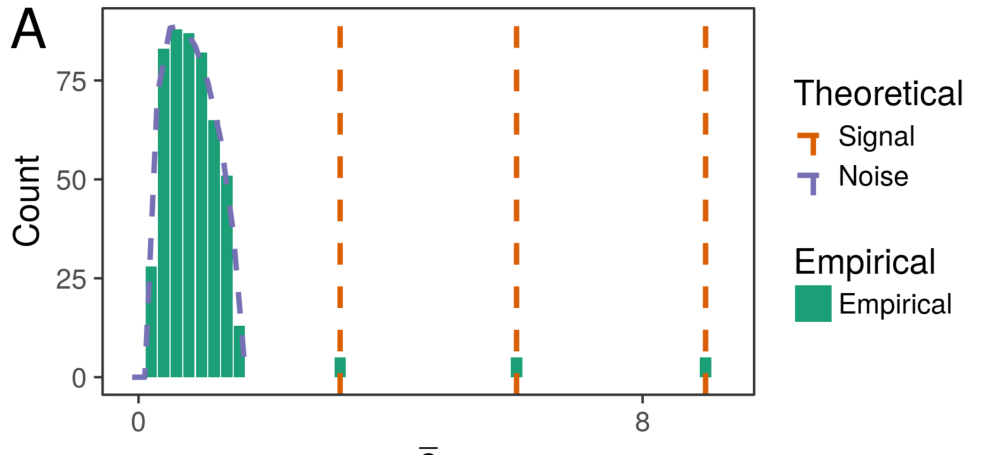
$$\overline{W} \equiv \overline{W}^{23}\overline{W}^{12}$$

$$\hat{y}^\mu = \overline{W}\hat{x}^\mu + z^\mu \longrightarrow$$

$$\Sigma^{11} = \sum_{\mu=1}^{P} \hat{x}^\mu \hat{x}^{\mu\,T} = I_{N_1 \times N_1}$$

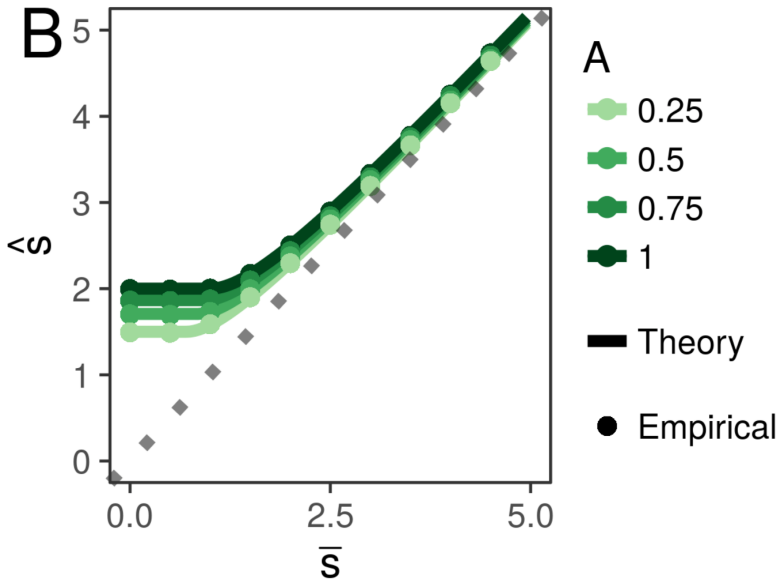
$$\Sigma^{31} = \sum_{\mu=1}^{P} \hat{y}^\mu \hat{x}^{\mu\,T} = \overline{W} + \tilde{Z}$$

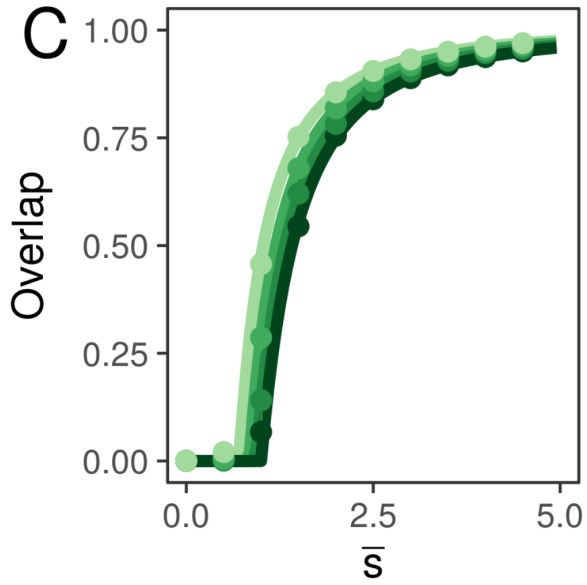
How the teacher is buried in the training data



$$\overline{W} = \sum_{\alpha=1}^{\overline{N_2}} \overline{s}^{\alpha}\, \overline{u}^{\alpha}\, \overline{v}^{\alpha\, T}$$

$$\Sigma^{31} = \sum_{\alpha=1}^{N_3} \hat{s}^{\alpha}\, \hat{u}^{\alpha}\, \hat{v}^{\alpha\, T}$$

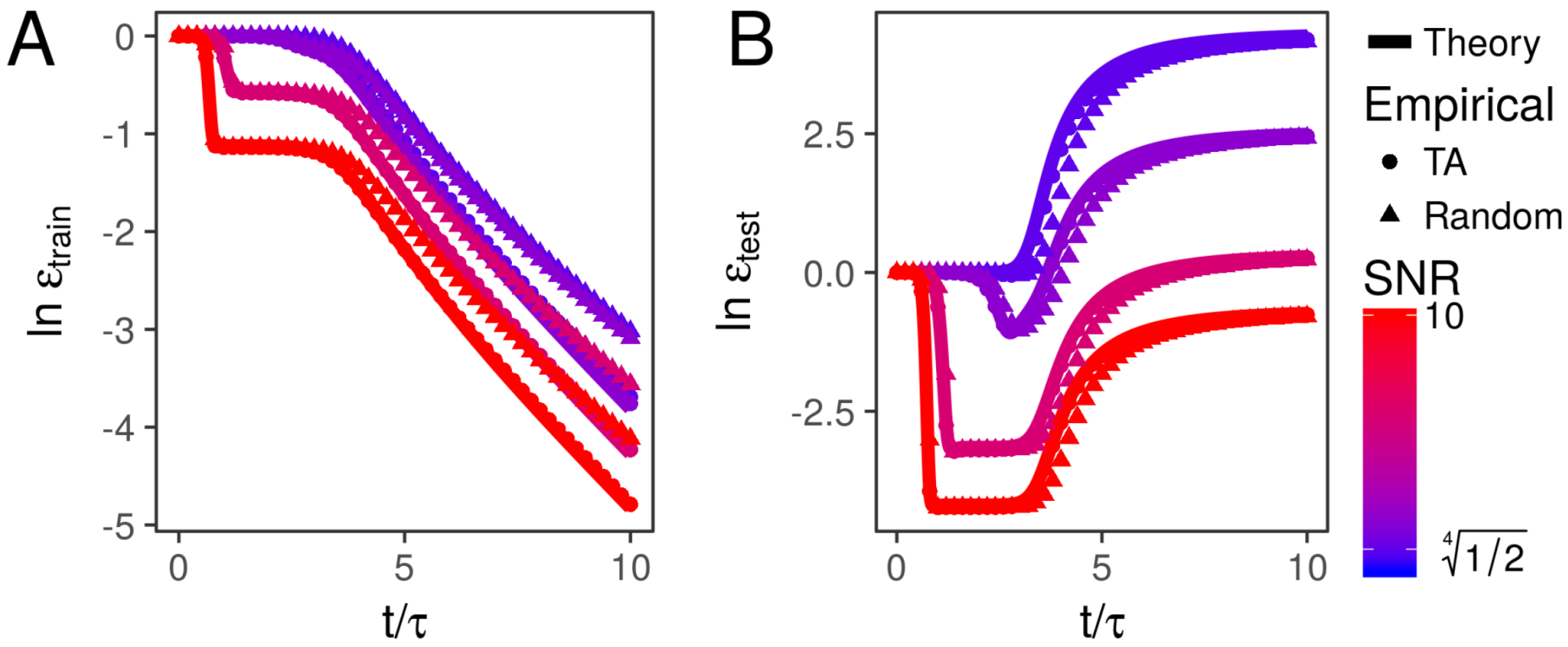




$\overline{s}$: teacher singular value

$\hat{s}$: training data singular value

Match between theory and numerics for training and testing error

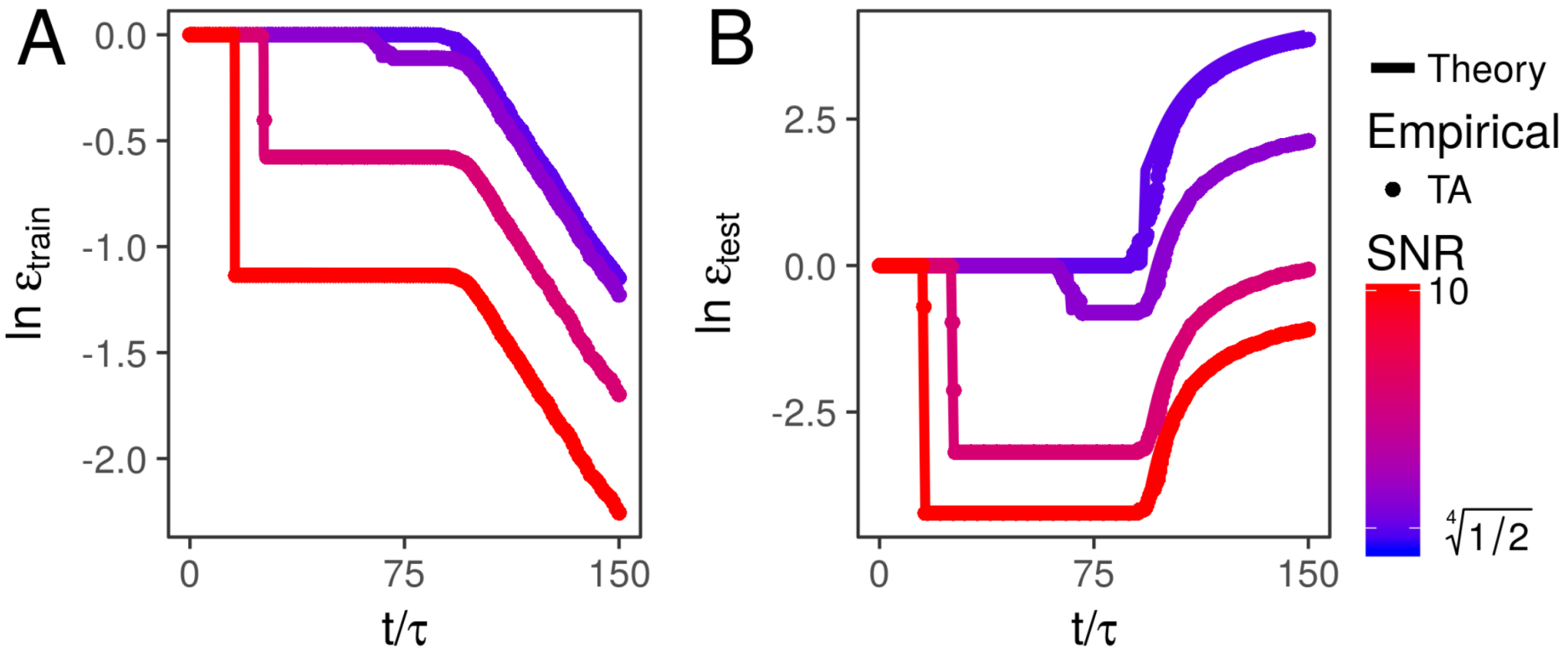


Rank N student, Rank 1 Teacher, both have one hidden layer.

Test error at the optimal early stopping time is independent of number of student hidden units!

It only depends on the structure of the data, not on the student architecture.

Match between theory and numerics for training and testing error



Rank N student, Rank 1 Teacher, student has 5 layers (3 hidden layers)

Test error at the optimal early stopping time is independent of number of student hidden units!

It only depends on the structure of the data, not on the student architecture.

At a coarse grained level: 3 puzzles of deep learning

Generalization: How can neural networks predict the response to new examples?

A. Saxe, J. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep neural networks, ICLR 2014.

A. Lampinen, J. McCelland, S. Ganguli, An analytic theory of generalization dynamics and transfer learning in deep linear networks, work in progress.

Expressivity: Why deep? What can a deep neural network "say" that a shallow network cannot?

B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, NIPS 2016.

Trainability: How can we optimize non-convex loss functions to achieve small training error?

Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice, J. Pennington, S. Schloenholz, and S. Ganguli, NIPS 2017.

The emergence of spectral universality in deep networks, J. Pennington, S. Schloenholz, and S. Ganguli, AISTATS 2018.

Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, NIPS 2014.

A theory of deep neural expressivity through transient input-output chaos

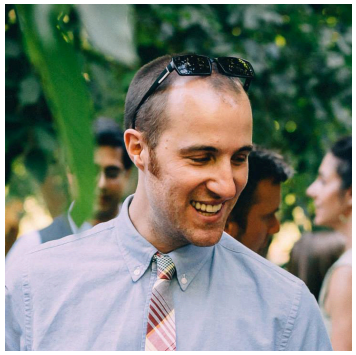
Stanford

Google

Ben Poole

Subhaneil Lahiri

Maithra Raghu



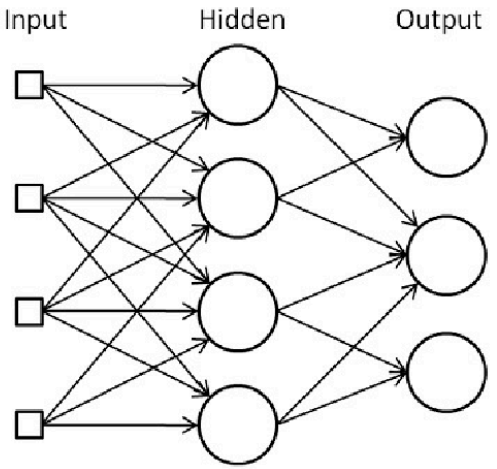
Jascha Sohl-Dickstein

Expressivity: what kinds of functions can a deep network express that shallow networks cannot?

Exponential expressivity in deep neural networks through transient chaos, B. Poole, S. Lahiri,M. Raghu, J. Sohl-Dickstein, S. Ganguli, NIPS 2016.

On the expressive power of deep neural networks, M.Raghu, B. Poole,J. Kleinberg, J. Sohl-Dickstein, S. Ganguli, under review, ICML 2017.

The problem of expressivity



Networks with one hidden layer are universal function approximators.

So why do we need depth?

Overall idea: there exist certain (special?) functions that can be computed:

a) efficiently using a deep network (poly # of neurons in input dimension)

b) but not by a shallow network (requires exponential # of neurons)

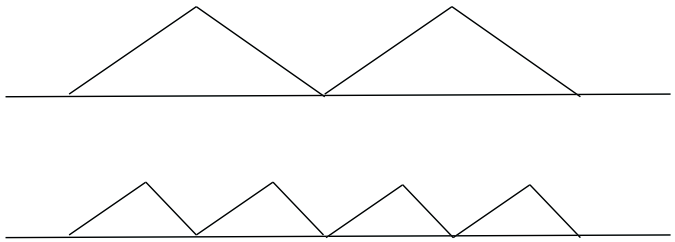
Intellectual traditions in boolean circuit theory: parity function is such a function for boolean circuits.

Seminal works on the expressive power of depth

Nonlinearity	Measure of Functional Complexity
Rectified Linear Unit (ReLu)	Number of linear regions

There exists a “saw-tooth” function computable by a deep network where the number of linear regions is exponential in the depth.

To approximate this function with a shallow network, one would require exponentially many more neurons.



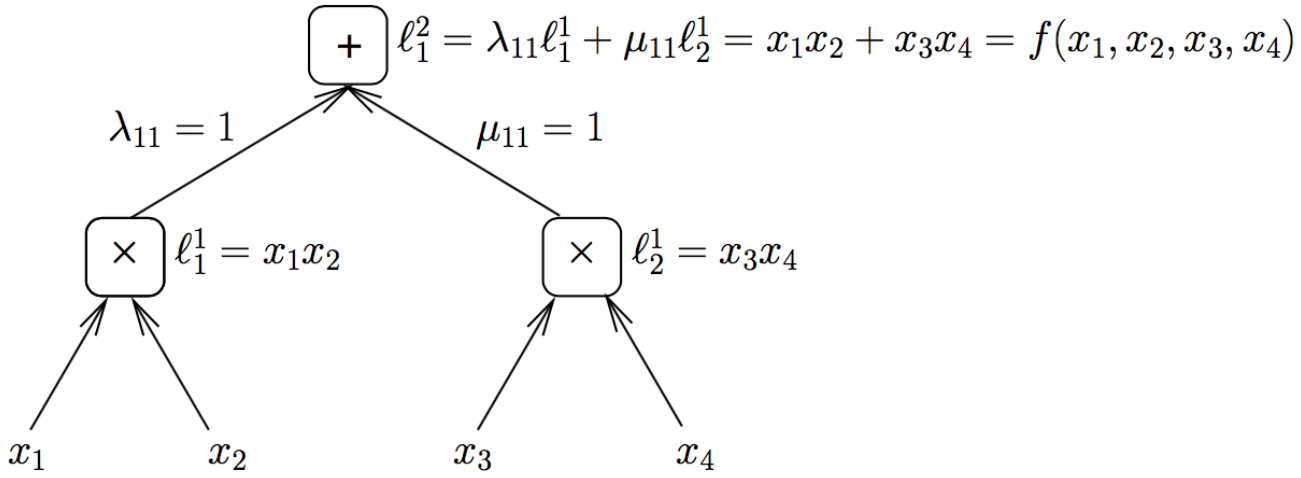
Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio.
On the number of linear regions of deep neural networks, NIPS 2014

Seminal works on the expressive power of depth

Nonlinearity	Measure of Functional Complexity
Sum-product network	Number of monomials

There exists a function computable by a deep network where the number of unique monomials is exponential in the depth.

To approximate this function with a shallow network, one would require exponentially many more neurons.



Olivier Delalleau and Yoshua Bengio. Shallow vs. deep sum-product networks, NIPS 2011.

Questions

The particular functions exhibited by prior work do not seem natural?

Are such functions rare curiosities?

Or is this phenomenon much more generic than these specific examples?

In some sense, is **any** function computed by a **generic** deep network
not efficiently computable by a shallow network?

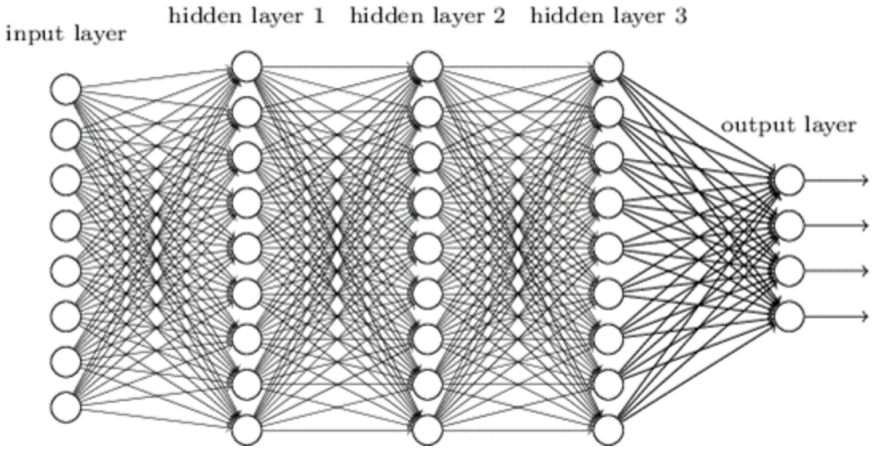
If so we would like a theory of deep neural expressivity that demonstrates this for

1) Arbitrary nonlinearities
2) A natural, general measure of functional complexity.

We will combine Riemannian geometry + dynamic mean field theory to show that even in generic, random deep neural networks, measures of functional curvature grow exponentially with depth but not width!

More over the origins of this exponential growth can be traced to chaos theory.

A maximum entropy ensemble of deep random networks



$$N_l = \text{number of neurons in layer l}$$

$$D = \text{depth}(l = 1, \ldots, D)$$

$$\mathbf{x}^l = \phi(\mathbf{h}^l)$$

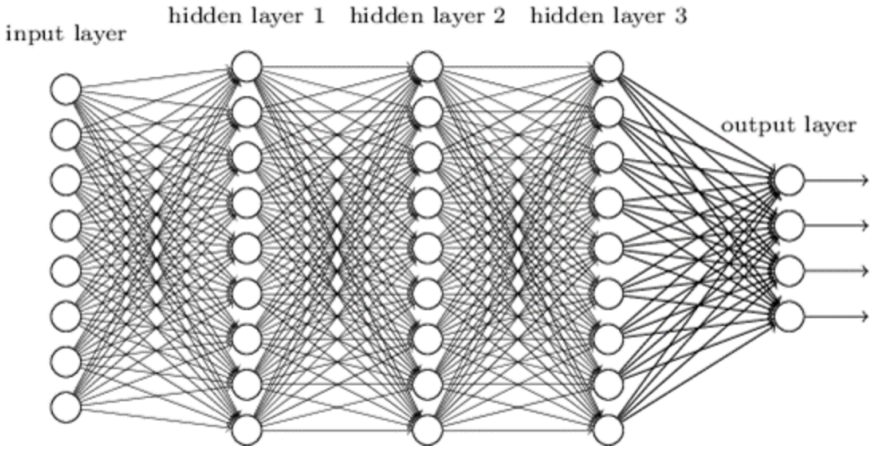
$$\mathbf{h}^l = \mathbf{W}^l \, \mathbf{x}^{l-1} + \mathbf{b}^l$$

Structure: i.i.d. random Gaussian weights and biases:

$$\mathbf{W}_{ij}^l \leftarrow \mathcal{N}\left(0, \frac{\sigma_w^2}{N^{l-1}}\right)$$

$$\mathbf{b}_i^l \leftarrow \mathcal{N}(0, \sigma_b^2)$$

Emergent, deterministic signal propagation in random neural networks



$$N_l = \text{number of neurons in layer l}$$

$$D = \text{depth}(l = 1, \ldots, D)$$

$$\mathbf{x}^l = \phi(\mathbf{h}^l)$$

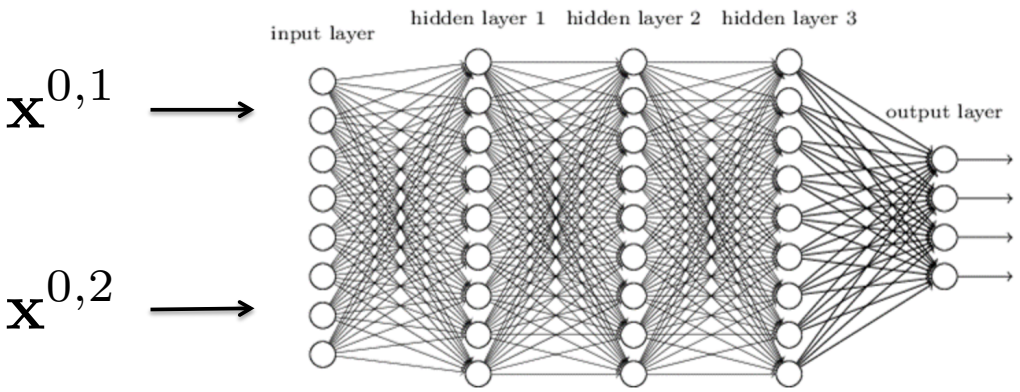
$$\mathbf{h}^l = \mathbf{W}^l \, \mathbf{x}^{l-1} + \mathbf{b}^l$$

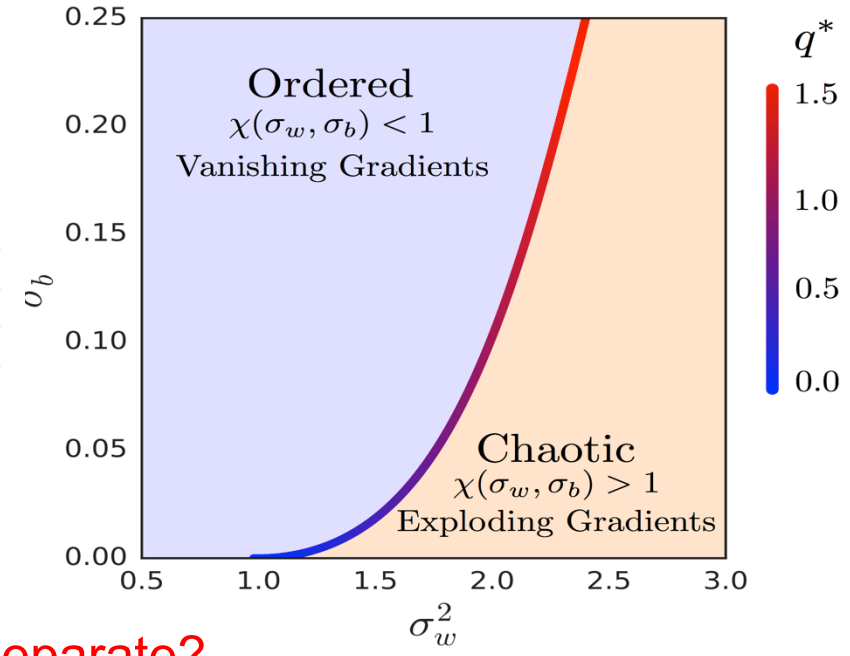
Question: how do simple input manifolds propagate through the layers?

A pair of points: Do they become more similar or more different, and how fast?

A smooth manifold: How does its curvature and volume change?

Propagation of two points through a deep network





Do nearby points come closer together or separate?

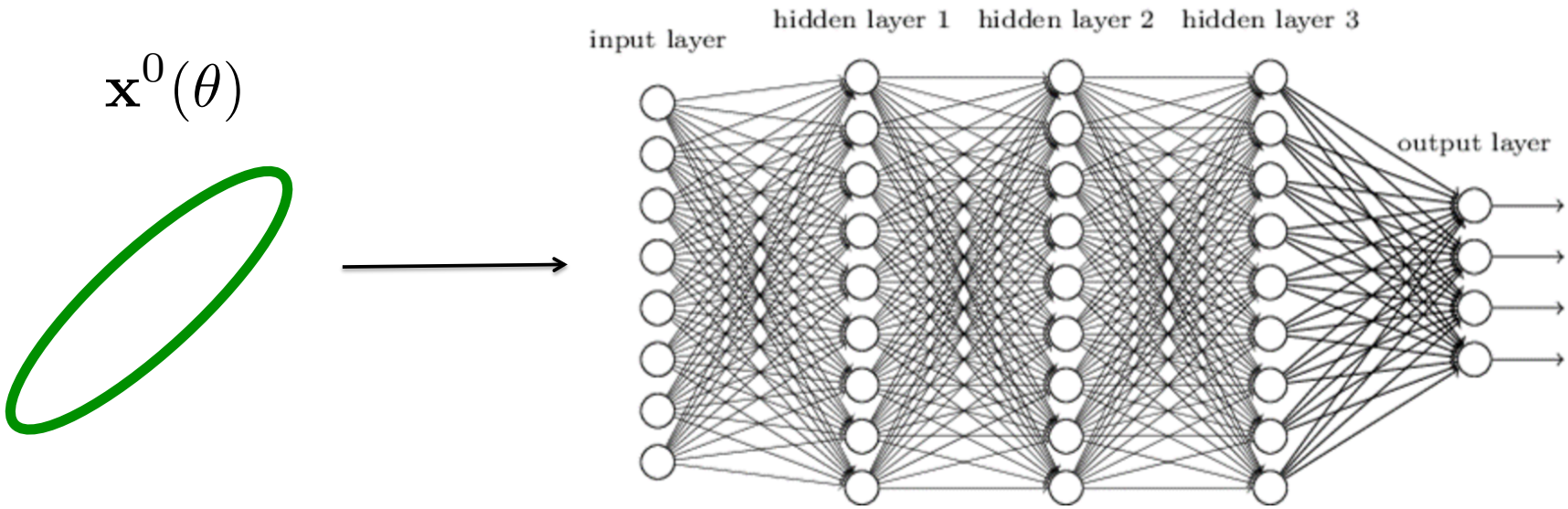
$$\chi = \frac{1}{N}\left\langle \mathrm{Tr}\left(\mathbf{DW}\right)^T \mathbf{DW}\right\rangle = \sigma_w^2 \int \mathcal{D}h \left[\phi'\left(\sqrt{q^*}h\right)\right]^2$$

χ is the mean squared singular value of the Jacobian across 1 layer

$\chi < 1$: nearby points come closer together; gradients exponentially vanish
$\chi > 1$: nearby points are driven apart; gradients exponentially explode

$$\mathbf{J} = \frac{\partial \mathbf{x}^L}{\partial \mathbf{h}^0} = \prod_{l=1}^{L} \mathbf{D}^l \mathbf{W}^l \qquad \frac{1}{N}\mathrm{Tr}\,\mathbf{J}^T\mathbf{J} = \chi^L$$

Propagation of a manifold through a deep network



The geometry of the manifold is captured by the similarity matrix - How similar two points are in internal representation space):

$$q^l(\theta_1, \theta_2) = \frac{1}{N_l} \sum_{i=1}^{N_l} \mathbf{h}_i^l[\mathbf{x}^0(\theta_1)]\, \mathbf{h}_i^l[\mathbf{x}^0(\theta_2)]$$

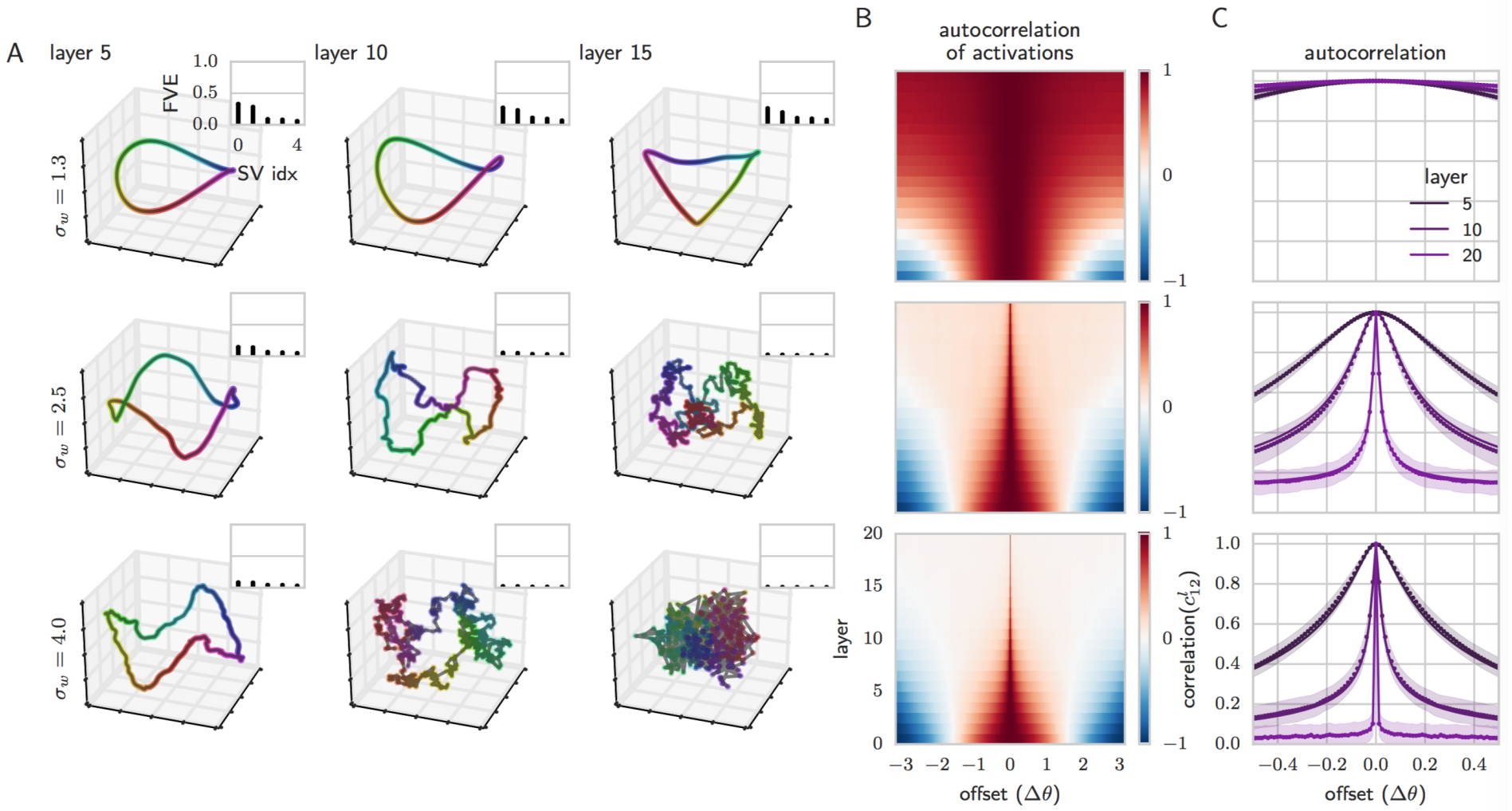
Or autocorrelation function:

$$q^l(\Delta\theta) = \int d\theta\, q^l(\theta, \theta + \Delta\theta)$$

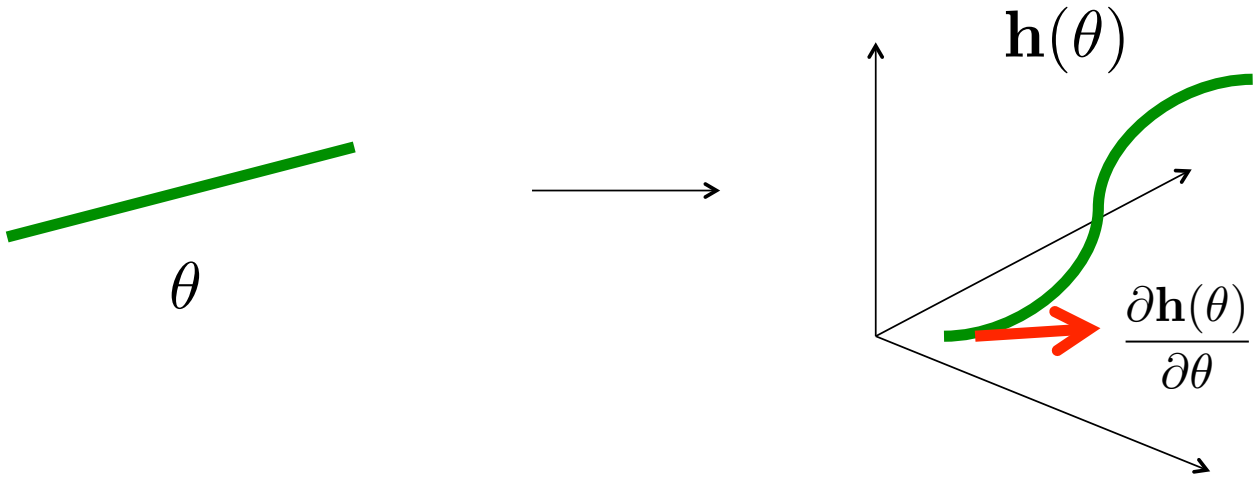
Propagation of a manifold through a deep network

$$\mathbf{h}^1(\theta) = \sqrt{N_1 q^*}\left[\mathbf{u}^0\cos(\theta) + \mathbf{u}^1\sin(\theta)\right]$$

A great circle
input manifold



Riemannian geometry I: Euclidean length



$$g^E(\theta) = \frac{\partial \mathbf{h}(\theta)}{\partial \theta} \cdot \frac{\partial \mathbf{h}(\theta)}{\partial \theta}$$

Metric on manifold coordinate θ induced by Euclidean metric in internal representation space **h**.

$$d\mathcal{L}^E = \sqrt{g^E(\theta)}d\theta$$

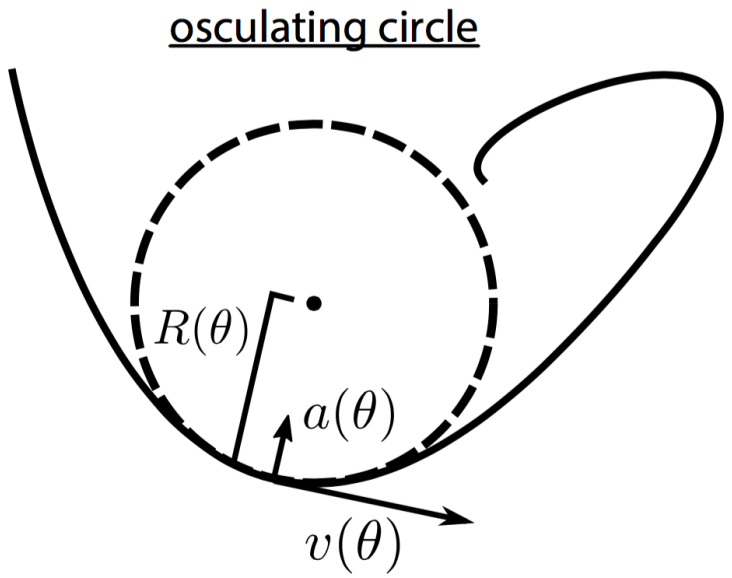
Length element: if one moves from Θ to Θ+ dΘ along the manifold, then one moves a distance dL^E in internal representation space

Riemannian geometry II: Extrinsic Gaussian Curvature

$\mathbf{h}(\theta)$ Point on the curve

$\mathbf{v}(\theta) = \dfrac{\partial \mathbf{h}(\theta)}{\partial \theta}$ Tangent or velocity vector

$\mathbf{a}(\theta) = \dfrac{\partial \mathbf{v}(\theta)}{\partial \theta}$ Acceleration vector



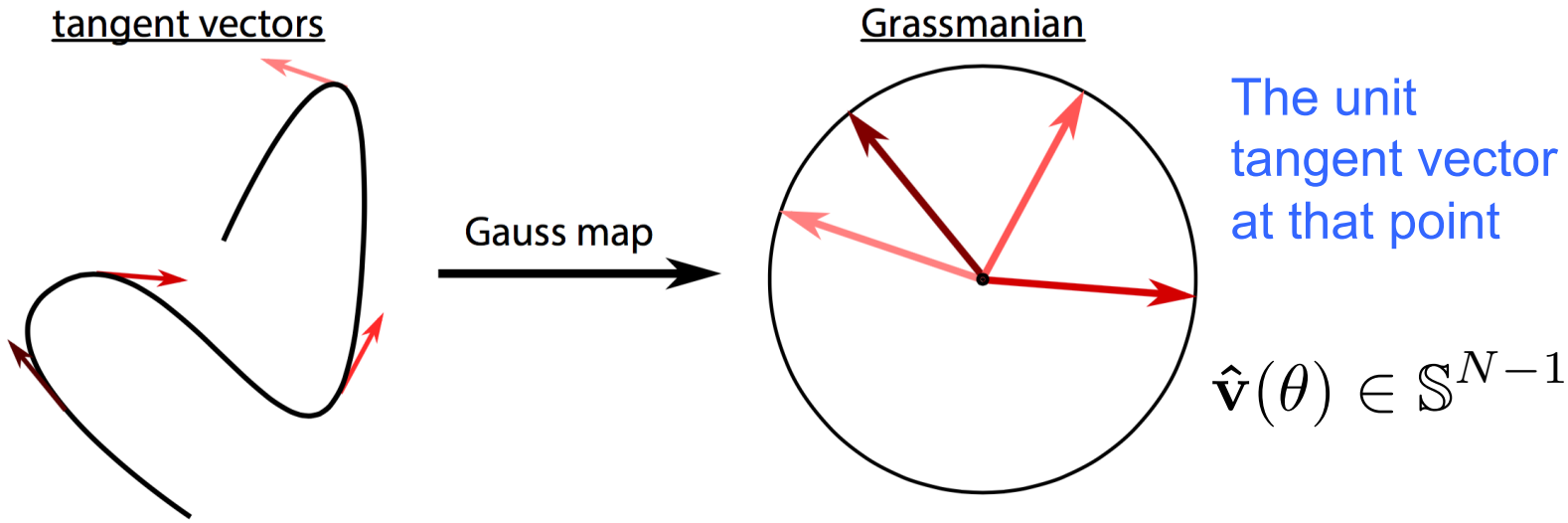
The velocity and acceleration vector span a 2 dimensional plane in N dim space.

Within this plane, there is a unique circle that touches the curve at $\mathbf{h}(\theta)$, with the same velocity and acceleration.

The Gaussian curvature $\kappa(\theta)$ is the inverse of the radius of this circle.

$$\kappa(\theta) = \sqrt{\frac{(\mathbf{v}\cdot\mathbf{v})(\mathbf{a}\cdot\mathbf{a}) - (\mathbf{v}\cdot\mathbf{a})^2}{(\mathbf{v}\cdot\mathbf{v})^3}}$$

Riemannian geometry III: The Gauss map and Grassmannian length



$$g^G(\theta) = \frac{\partial \hat{\mathbf{v}}(\theta)}{\partial \theta} \cdot \frac{\partial \hat{\mathbf{v}}(\theta)}{\partial \theta}$$

Metric on manifold coordinate θ induced by metric on the Grassmannian: how quickly unit tangent vector changes

$$d\mathcal{L}^G = \sqrt{g^G(\theta)}d\theta$$

Length element: if one moves from Θ to Θ+ dΘ along the manifold, then one moves a distance dL^G Along the Grassmanian

$$g^G(\theta) = \kappa(\theta)^2 g^E(\theta)$$

Grassmannian length, Gaussian curvature and Euclidean length

An example: the great circle

$$\mathbf{h}^1(\theta) = \sqrt{Nq}\left[\mathbf{u}^0\cos(\theta) + \mathbf{u}^1\sin(\theta)\right]$$

A great circle input manifold

Euclidean length

$$g^E(\theta) = Nq$$

$$\mathcal{L}^E = 2\pi\sqrt{Nq}$$

Gaussian Curvature

$$\kappa(\theta) = 1/\sqrt{Nq}$$

Grassmannian Length

$$g^G(\theta) = 1$$

$$\mathcal{L}^G = 2\pi$$

Behavior under isotropic linear expansion via multiplicative stretch χ_1:

$$\mathcal{L}^E \to \sqrt{\chi_1}\,\mathcal{L}^E$$

$$\kappa \to \frac{1}{\sqrt{\chi_1}}\,\kappa$$

$$\mathcal{L}^G \to \mathcal{L}^G$$

$\chi_1 < 1$	Contraction	Increase	Constant
$\chi_1 > 1$	Expansion	Decrease	Constant

An example: the great circle

$$\mathbf{h}^1(\theta) = \sqrt{Nq}\left[\mathbf{u}^0\cos(\theta) + \mathbf{u}^1\sin(\theta)\right]$$

A great circle input manifold

Euclidean length	Gaussian Curvature	Grassmannian Length
$g^E(\theta) = Nq$	$\kappa(\theta) = 1/\sqrt{Nq}$	$g^G(\theta) = 1$
$\mathcal{L}^E = 2\pi\sqrt{Nq}$		$\mathcal{L}^G = 2\pi$

Behavior under isotropic linear expansion via multiplicative stretch χ_1:

	$\mathcal{L}^E \to \sqrt{\chi_1}\,\mathcal{L}^E$	$\kappa \to \frac{1}{\sqrt{\chi_1}}\,\kappa$	$\mathcal{L}^G \to \mathcal{L}^G$
$\chi_1 < 1$	Contraction	Increase	Constant
$\chi_1 > 1$	Expansion	Decrease	Constant

Theory of curvature propagation in deep networks

$$\bar{g}^{E,l} = \chi_1 \, \bar{g}^{E,l-1} \qquad \bar{g}^{E,1} = q^*$$

$$(\bar{\kappa}^l)^2 = 3\frac{\chi_2}{\chi_1^2} + \frac{1}{\chi_1}(\bar{\kappa}^{l-1})^2 \qquad (\bar{\kappa}^1)^2 = \frac{1}{q^*}$$

$$\chi_1 = \sigma_w^2 \int \mathcal{D}z \left[\phi'\left(\sqrt{q^*}z\right)\right]^2$$

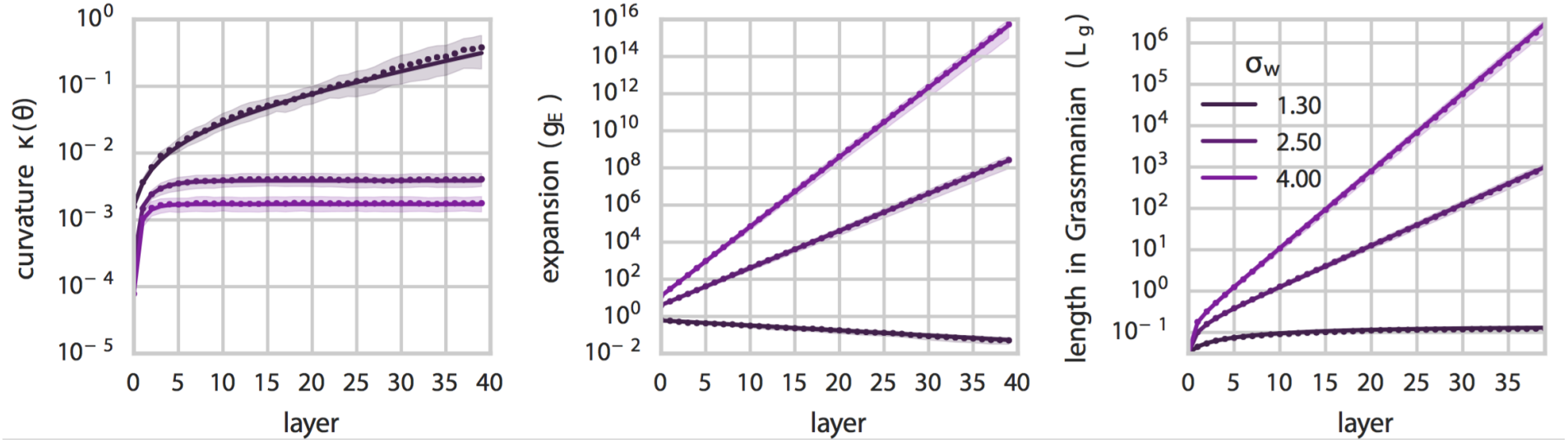
$$\chi_2 = \sigma_w^2 \int \mathcal{D}z \left[\phi''\left(\sqrt{q^*}z\right)\right]^2$$

Modification of existing curvature due to stretch

Addition of new curvature due to nonlinearity

		Local Stretch	Extrinsic Curvature	Grassmannian Length
Ordered:	$\chi_1 < 1$	Contraction	Explosion	Constant
Chaotic:	$\chi_1 > 1$	Expansion	Attenuation + Addition	Exponential Growth

Curvature propagation: theory and experiment

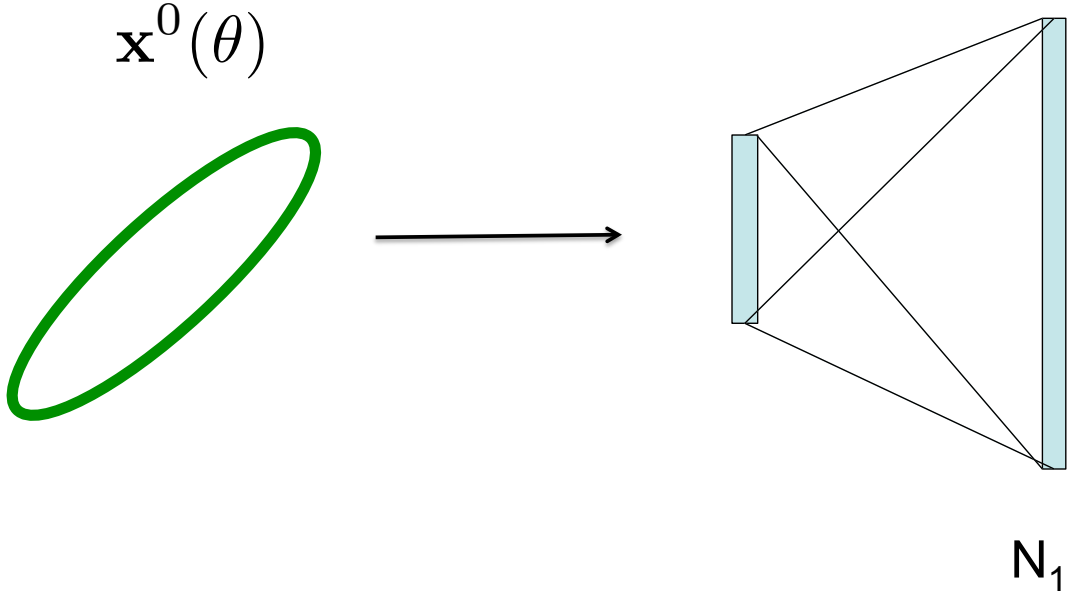


Unlike linear expansion, deep neural signal propagation can:

1) exponentially expand length,
2) without diluting Gaussian curvature,
3) thereby yielding exponential growth of Grassmannian length.

As a result, the circle will become fill space as it winds around at a constant rate of curvature to explore many dimensions!

Exponential expressivity is not achievable by shallow nets



Consider a shallow network with 1 hidden layer $\mathbf{x}^1$, one input layer $\mathbf{x}^0$, with $\mathbf{x}^1 = \phi(\mathbf{W}^1\mathbf{x}^0) + \mathbf{b}^1$, and a linear readout layer. How complex can the hidden representation be as a function of its width N_1, relative to the results above for depth? We prove a general upper bound on $\mathcal{L}^E$ (see SM):

Theorem 1. *Suppose $\phi(h)$ is monotonically non-decreasing with bounded dynamic range R, i.e. $\max_h \phi(h) - \min_h \phi(h) = R$. Further suppose that $\mathbf{x}^0(\theta)$ is a curve in input space such that no 1D projection of $\partial_\theta \mathbf{x}(\theta)$ changes sign more than s times over the range of θ. Then for any choice of $\mathbf{W}^1$ and $\mathbf{b}^1$ the Euclidean length of $\mathbf{x}^1(\theta)$, satisfies $\mathcal{L}^E \leq N_1(1+s)R$.*

Summary

We have combined Riemannian geometry with dynamical mean field theory to study the emergent deterministic properties of signal propagation in deep nonlinear nets.

We derived analytic recursion relations for Euclidean length, correlations, curvature, and Grassmannian length as simple input manifolds propagate forward through the network.

We obtain an excellent quantitative match between theory and simulations.

Our results reveal the existence of a transient chaotic phase in which the network expands input manifolds without straightening them out, leading to "space filling" curves that explore many dimensions while turning at a constant rate. The number of turns grows exponentially with depth.

Such exponential growth does not happen with width in a shallow net.

Chaotic deep random networks can also take exponentially curved N-1 Dimensional decision boundaries in the input and flatten them into Hyperplane decision boundaries in the final layer: exponential disentangling!

At a coarse grained level: 3 puzzles of deep learning

Generalization: How can neural networks predict the response to new examples?

A. Saxe, J. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep neural networks, ICLR 2014.

A. Lampinen, J. McCelland, S. Ganguli, An analytic theory of generalization dynamics and transfer learning in deep linear networks, work in progress.

Expressivity: Why deep? What can a deep neural network "say" that a shallow network cannot?

B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, NIPS 2016.

Trainability: How can we optimize non-convex loss functions to achieve small training error?

Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice, J. Pennington, S. Schloenholz, and S. Ganguli, NIPS 2017.

The emergence of spectral universality in deep networks, J. Pennington, S. Schloenholz, and S. Ganguli, AISTATS 2018.

Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, NIPS 2014.

Beyond manifold geometry to entire Jacobian singular value distributions

Andrew Saxe
Harvard

Sam Schoenholz
Google Brain

Jeff Pennington
Google Brain

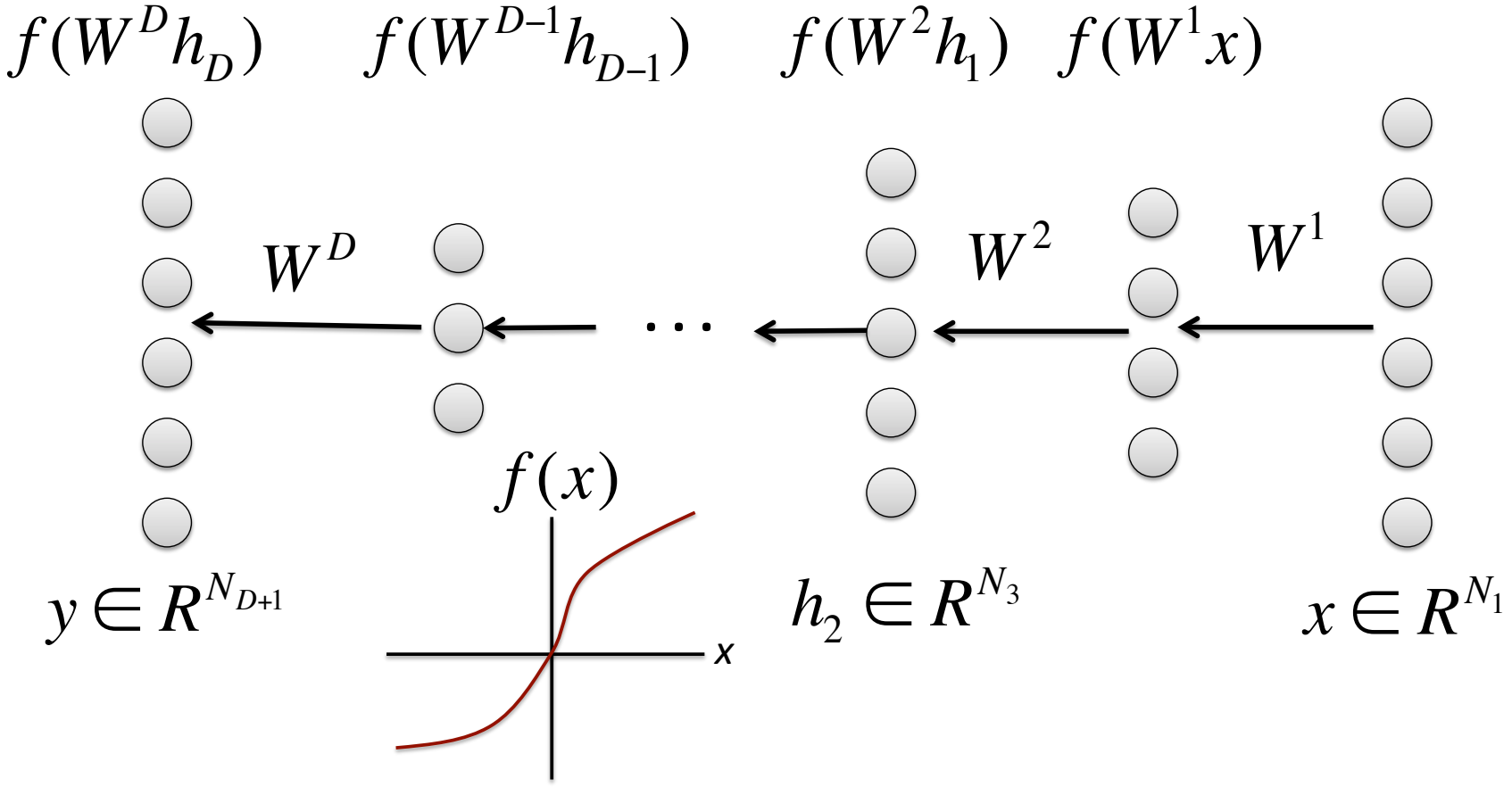
Question: how do random initializations and nonlinearities impact learning dynamics?

Exact solutions to the nonlinear dynamics of learning in deep linear networks, A. Saxe, J. McClelland, S. Ganguli, ICLR 2014.

Investigating the learning dynamics of deep neural networks using random matrix theory, J. Pennington, S. Schoenholz, S. Ganguli, ICML 2017.

The emergence of spectral universality in deep networks, J. Pennington, S. Schloenholz, and S. Ganguli, AISTATS 2018.

A Deep network



End to end Jacobian:

$$J = F^D W^D \ldots F^3 W^3 F^2 W^2 F^2 W^1$$

Prediction from an analytic theory of nonlinear learning dynamics in deep linear nets:

If you initialize with random orthogonal weights (or rectangular matrices with random singular vectors but all singular values = 1) then:

Learning time, in number of epochs, will be independent of depth even as the depth goes to infinity.

If you initialize with random Gaussian weights then:

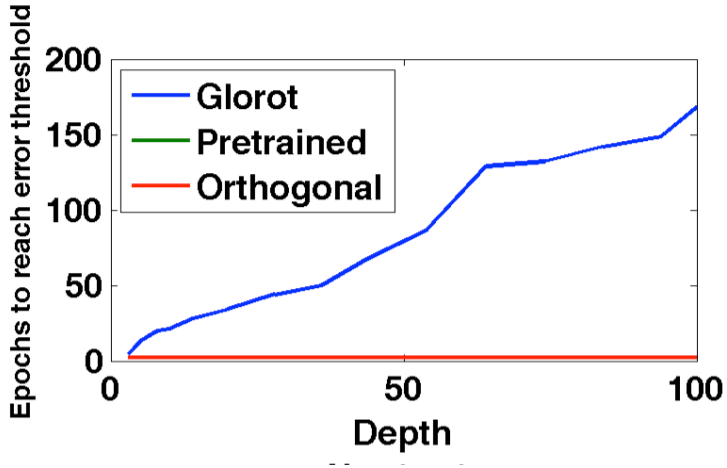
Learning time will grow with depth, and cannot remain constant.

Exact solutions to the nonlinear dynamics of learning in deep linear networks, A. Saxe, J. McClelland, S. Ganguli, ICLR 2014.

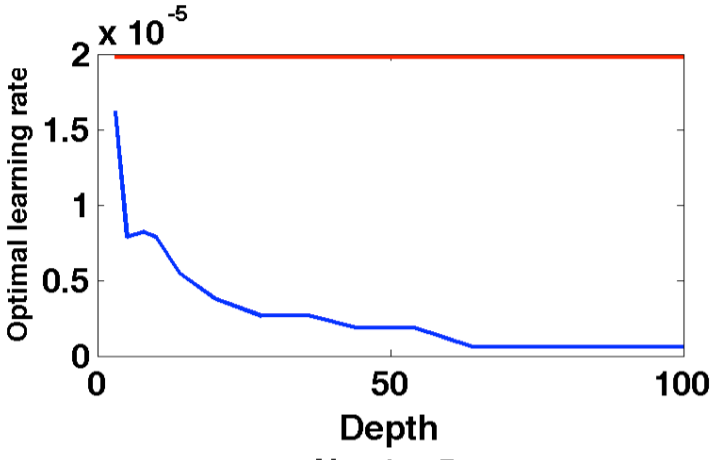
Theoretical prediction verified: Depth independent training times

- Deep *linear* networks on MNIST
- Scaled random Gaussian initialization (Glorot & Bengio, 2010)

Time to criterion



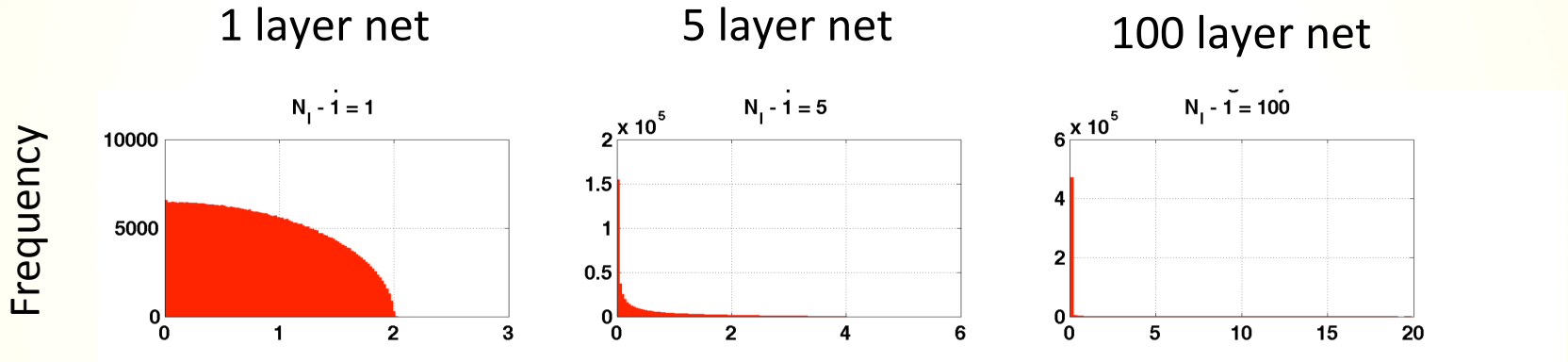
Optimal learning rate



- Pretrained and orthogonal have fast **depth-independent** training times!

Random vs orthogonal

- Gaussian preserves norm of random vector *on average*

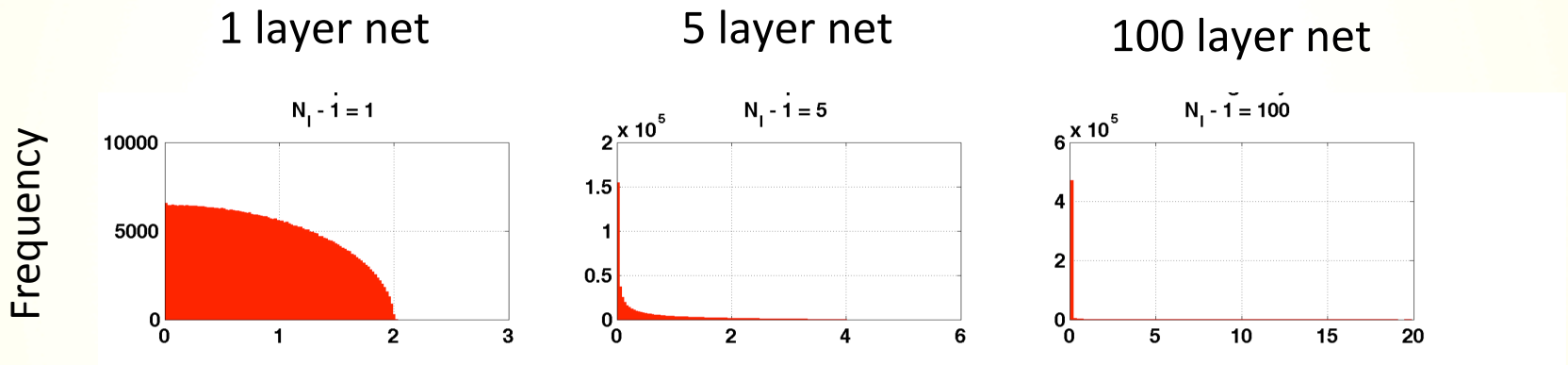


Singular values of $W^{tot} = \prod_{i=1}^{N_l - 1} W^i$

- *Attenuates* on subspace of high dimension
- *Amplifies* on subspace of low dimension

Random vs orthogonal

- Glorot preserves norm of random vector *on average*

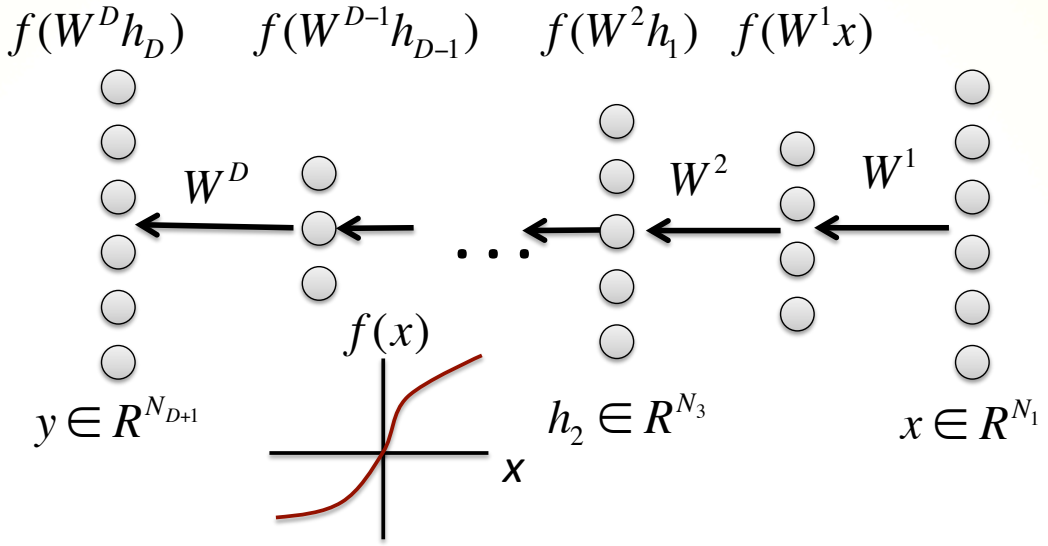


Singular values of $W^{tot} = \prod_{i=1}^{N_l - 1} W^i$

- Orthogonal preserves norm of all vectors *exactly*

All singular values of $W^{tot} = 1$

Analysis of gradient flow in nonlinear deep nets through free probability



End to end Jacobian:

$$J = F^D W^D \ldots F^3 W^3 F^2 W^2 F^2 W^1$$

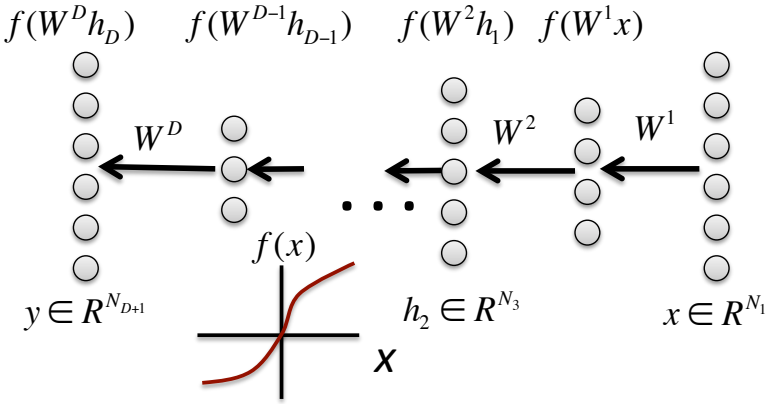
Free probability theory for random matrix products:

$$[\, S_F(z)\, S_W(z)\,]^D$$

$\sigma(A)$ $\sigma(B)$ $\sigma(AB)$

$S_A(z)$ $S_B(z)$ $\longrightarrow$ $S_A(z)\, S_B(z)$

Beyond mean squared singular value: free probability analysis of all the Jacobian singular values



The end to end Jacobian is a product of random matrices.

$$\mathbf{J} = \frac{\partial \mathbf{x}^L}{\partial \mathbf{h}^0} = \prod_{l=1}^{L} \mathbf{D}^l \mathbf{W}^l \qquad \frac{1}{N} \mathrm{Tr}\, \mathbf{J}^T \mathbf{J} = \chi^L$$

If **A** and **B** are **freely independent,** then the spectrum of the product **AB** can be computed using the S-transform:

$$\begin{matrix} \rho_A(\lambda) \to G_A(z) \to S_A \\ \rho_B(\lambda) \to G_B(z) \to S_B \end{matrix} \searrow \nearrow \; S_A S_B = S_{AB} \to G_{AB}(z) \to \rho_{AB}(\lambda)$$

Free probability analysis of Jacobian singular values

Scaling properties as the depth D goes to infinity:

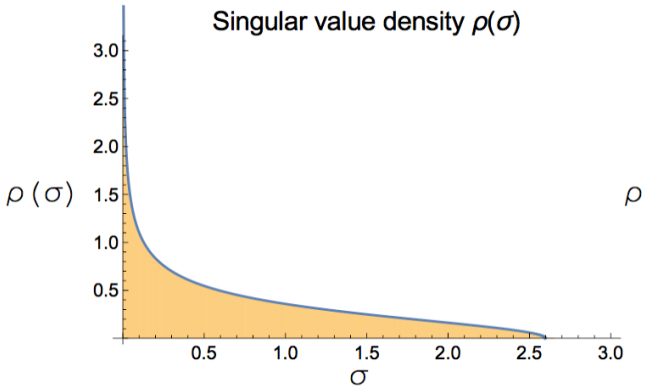
	Fraction of singular values Within 1-ε to 1+ε		Maximum singular value	
	Gaussian		Orthogonal	
Linear:	1/D	D	1	1
ReLu:	1/D	D	1/D	D
Tanh	1/D	D	**O(1)**	**O(1)**

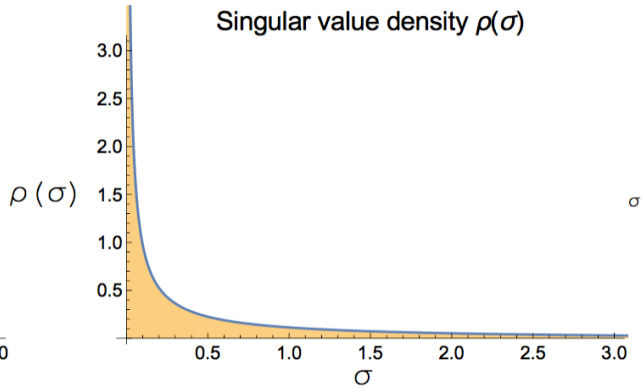
Free probability analysis of Jacobian singular values

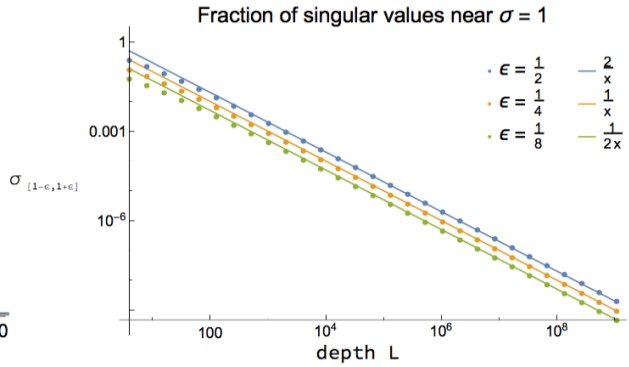
Example: linear network Gaussian weights at critical gain = 1

D = 2

D = 10





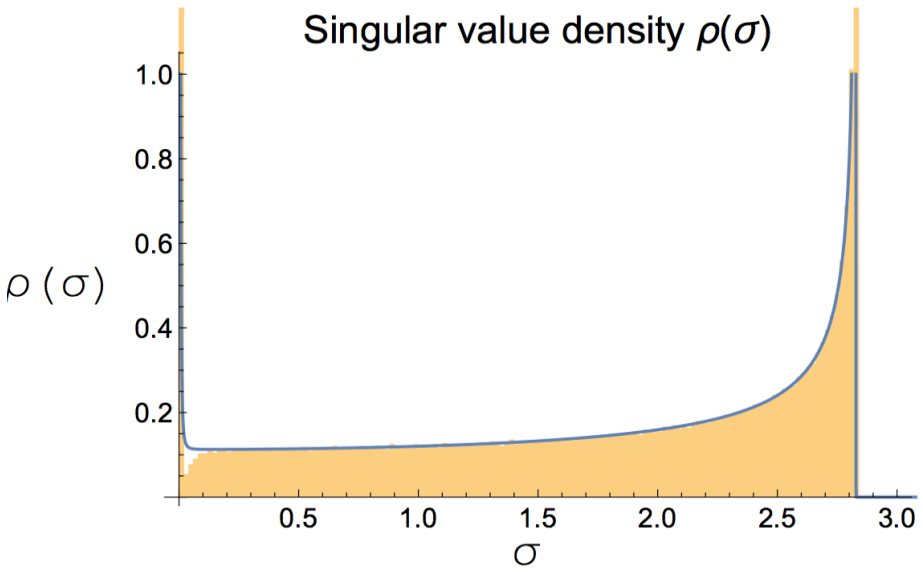


As depth D increases, the tail spreads out over an extent O(D)
and the “middle” around 1 falls off as O(1/D)

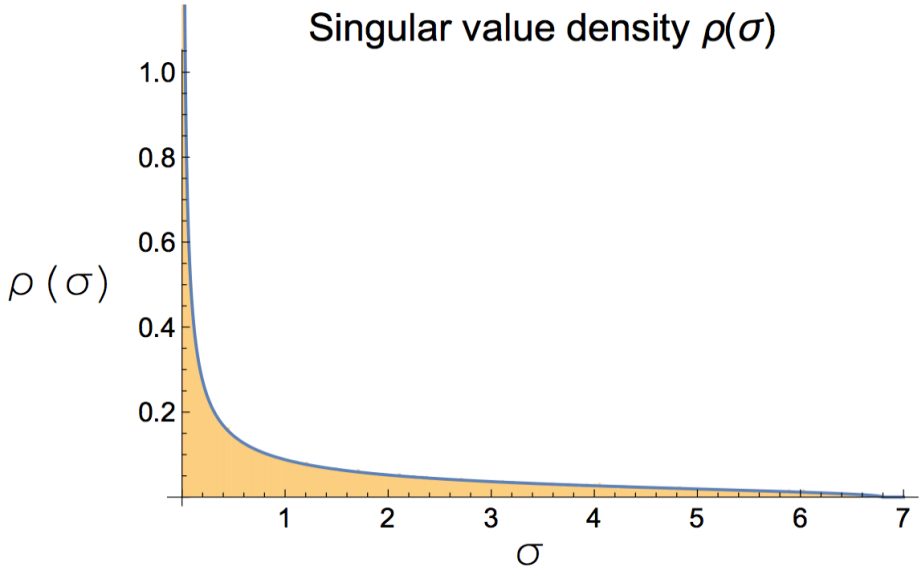
Free probability analysis of Jacobian singular values

Example: ReLU network with orthogonal weights at critical gain = $2^{1/2}$

D = 3



D = 10



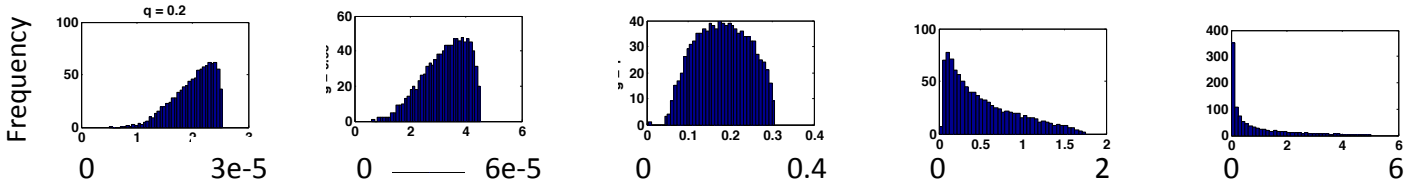
As depth D increases, the tail spreads out over an extent O(D)
and the “middle” around 1 falls off as O(1/D)

Free probability analysis of Jacobian singular values

Example: Tanh network with orthogonal weights at critical gain = 1

D = 100

Singular values of J



Gain

g=0.9 *g=0.95* *g=1* *g=1.05* *g=1.1*

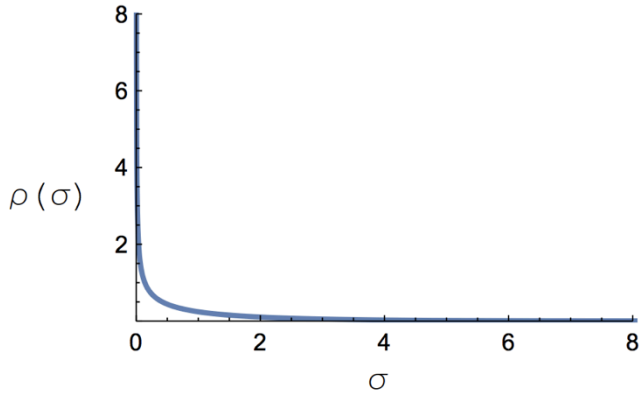
As depth D increases, the entire singular value distribution stabilizes to a well defined limit distribution!!

There is no extending tail that grows with depth!

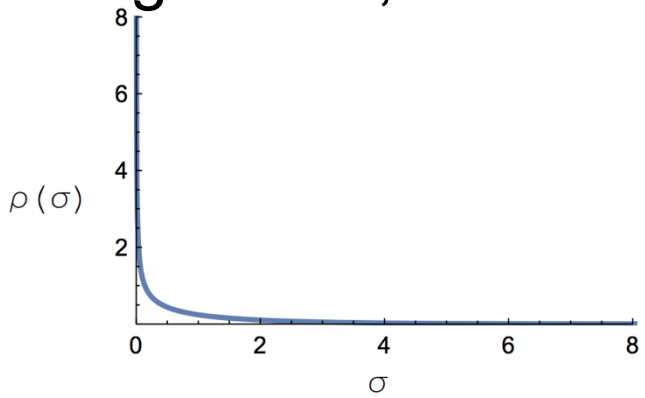
The emergence of spectral universality in deep networks,
J. Pennington, S. Schloenholz, and S. Ganguli, AISTATS 2018.

Free probability analysis of Jacobian singular values

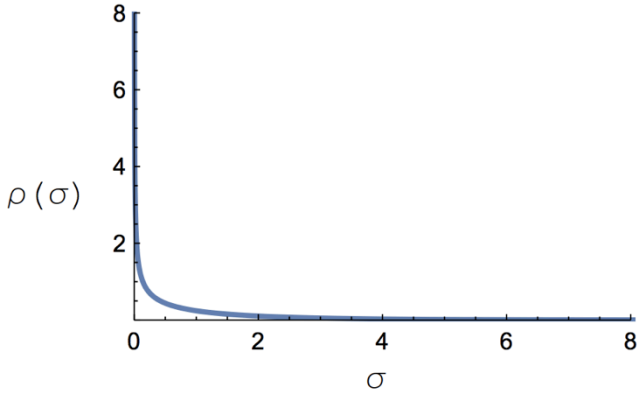
Gaussian W, any f



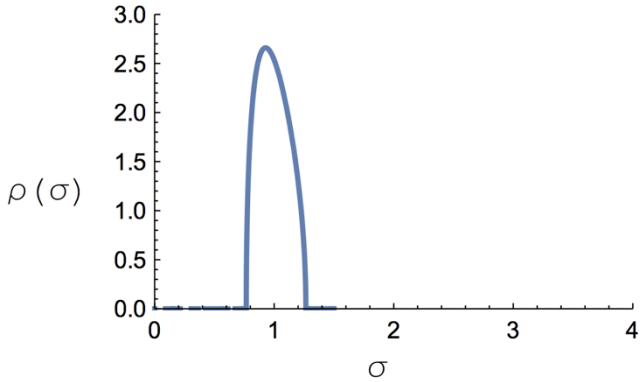
Orthogonal W, ReLU



Orthogonal W, tanh, $\sigma_w \gg 1$



Orthogonal W, tanh, $\sigma_w \sim 1+1/L$



Theorem: For Gaussian weights no nonlinearity can achieve dynamical isometry

Theorem: For ReLU, no random weights can achieve dynamical isometry

Free probability analysis of Jacobian singular values

Scaling properties as the depth D goes to infinity:

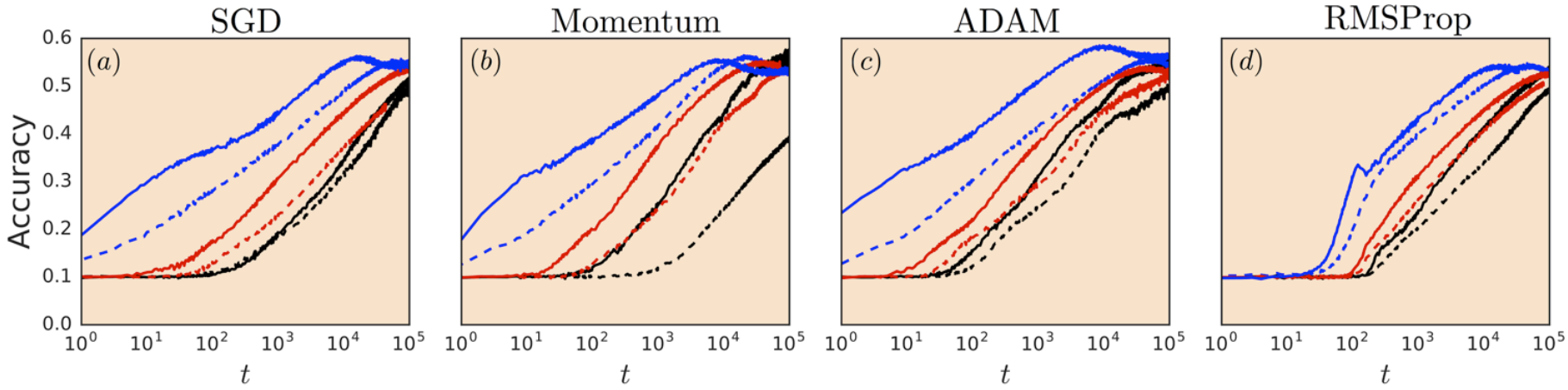
	Fraction of singular values Within 1-ε to 1+e		Maximum singular value	
	Gaussian		Orthogonal	
Linear:	1/L	L	1	1
ReLu:	1/L	L	1/L	L
Tanh	1/L	L	**O(1)**	**O(1)**

Theoretical prediction:

Sigmoidal networks with orthogonal weights can learn faster than ReLU networks with orthogonal weights

Learning speed: with orthogonal weights, Sigmoidal can outperform ReLu

Test accuracy on CIFAR-10



Tanh Orthogonal, $\sigma_w^2 = 1.05$

Tanh Orthogonal, $\sigma_w^2 = 2$

Relu Orthogonal, $\sigma_w^2 = 2$

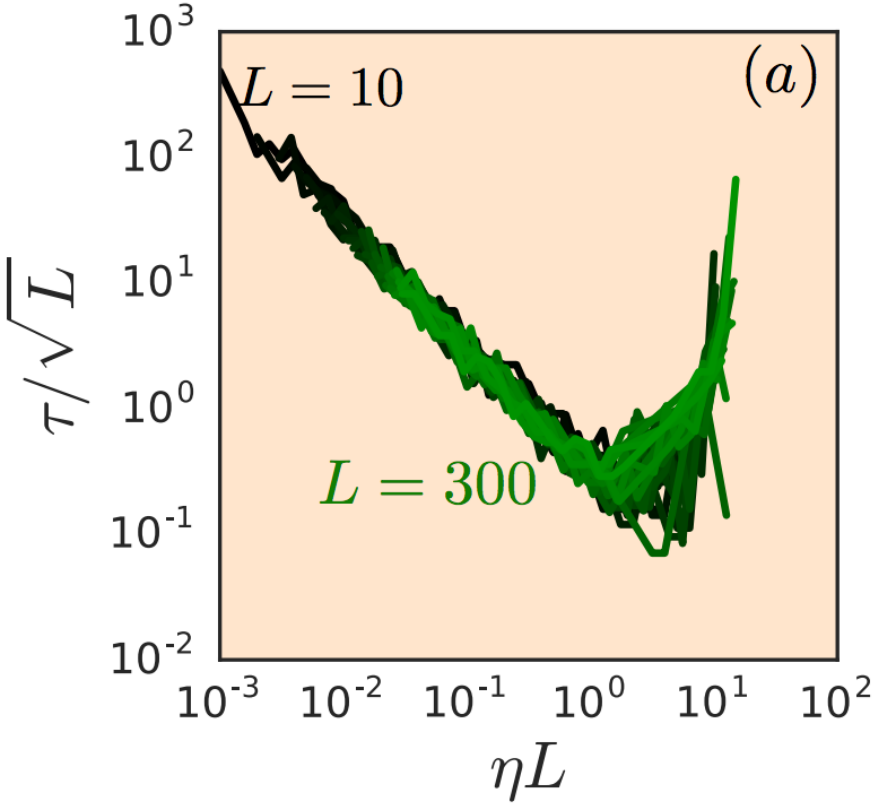
Tanh Gaussian, $\sigma_w^2 = 1.05$

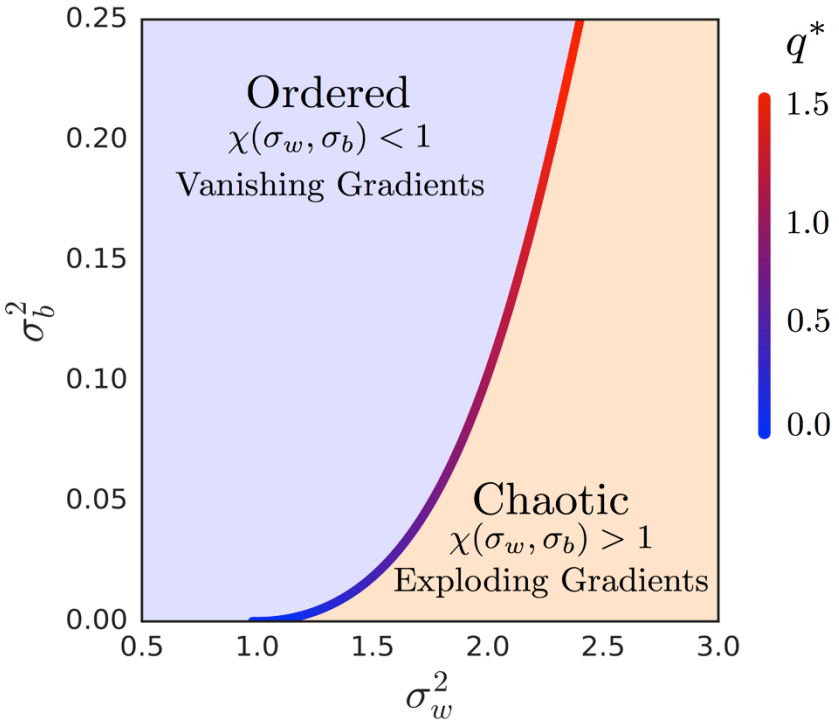
Tanh Gaussian, $\sigma_w^2 = 2$

Relu Gaussian, $\sigma_w^2 = 2$

Training time is sublinear in depth

A new scaling relation governing learning time as a function of depth





Optimal learning rate ~ 1/depth
training time ~ sqrt(depth)

Summary

An order to chaos phase transition governs the dynamics of random deep networks, often used for initialization.

Not all networks at the edge of chaos - with neither vanishing nor exploding gradients are created equal.

The entire Jacobian singular value distribution, and not just its second moment impacts learning speed.

We used introduced free probability theory to deep learning to compute this entire distribution.

We found tanh networks with orthogonal weights have well conditioned Jacobians, but ReLU networks with orthogonal weights, or any network with Gaussian weights does not.

Correspondingly, we found that with orthogonal weights, tanh networks learn Faster than ReLU networks.

Controlling the entire singular value distribution at initialization may be an important architectural design principle in deep learning.

References

- M. Advani and S. Ganguli, An equivalence between high dimensional Bayes optimal inference and M-estimation, NIPS 2016.
- M. Advani and S. Ganguli, Statistical mechanics of optimal convex inference in high dimensions, Physical Review X, 6, 031034, 2016.
- A. Saxe, J. McClelland, S. Ganguli, Learning hierarchical category structure in deep neural networks, Proc. of the 35th Cognitive Science Society, pp. 1271-1276, 2013.
- A. Saxe, J. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep neural networks, ICLR 2014.
- Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, NIPS 2014.
- B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, NIPS 2016.
- S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, Deep information propagation, https://arxiv.org/abs/1611.01232, under review at ICLR 2017.
- S. Lahiri, J. Sohl-Dickstein and S. Ganguli, A universal tradeoff between energy speed and accuracy in physical communication, arxiv 1603.07758
- A memory frontier for complex synapses, S. Lahiri and S. Ganguli, NIPS 2013.
- Continual learning through synaptic intelligence, F. Zenke, B. Poole, S. Ganguli, ICML 2017.
- Modelling arbitrary probability distributions using non-equilibrium thermodynamics, J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, ICML 2015.
- Deep Knowledge Tracing, C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas, J. Sohl-Dickstein, NIPS 2015.
- Deep learning models of the retinal response to natural scenes, L. McIntosh, N. Maheswaranathan, S. Ganguli, S. Baccus, NIPS 2016.
- Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice, J. Pennington, S. Schloenholz, and S. Ganguli, NIPS 2017.
- Variational walkback: learning a transition operator as a recurrent stochastic neural net, A. Goyal, N.R. Ke, S. Ganguli, Y. Bengio, NIPS 2017.
- The emergence of spectral universality in deep networks, J. Pennington, S. Schloenholz, and S. Ganguli, AISTATS 2018.