Statistical mechanics of deep learning

Surya Ganguli

Dept. of Applied Physics,
Neurobiology,
and Electrical Engineering

Stanford University

Funding:

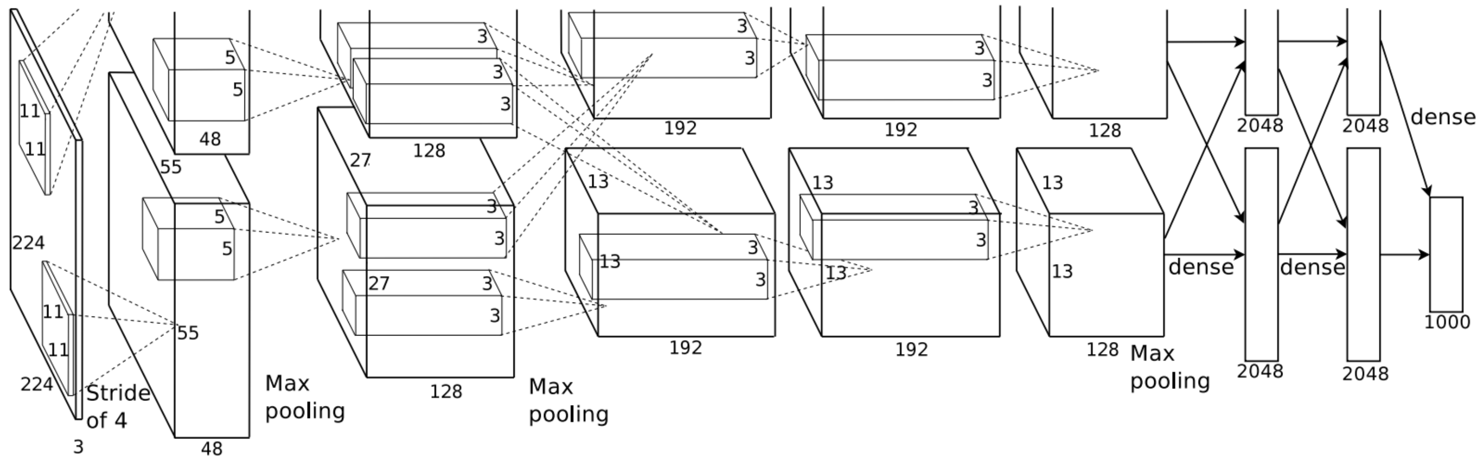
Bio-X Neuroventures
Burroughs Wellcome
Genentech Foundation
James S. McDonnell Foundation
McKnight Foundation
National Science Foundation

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

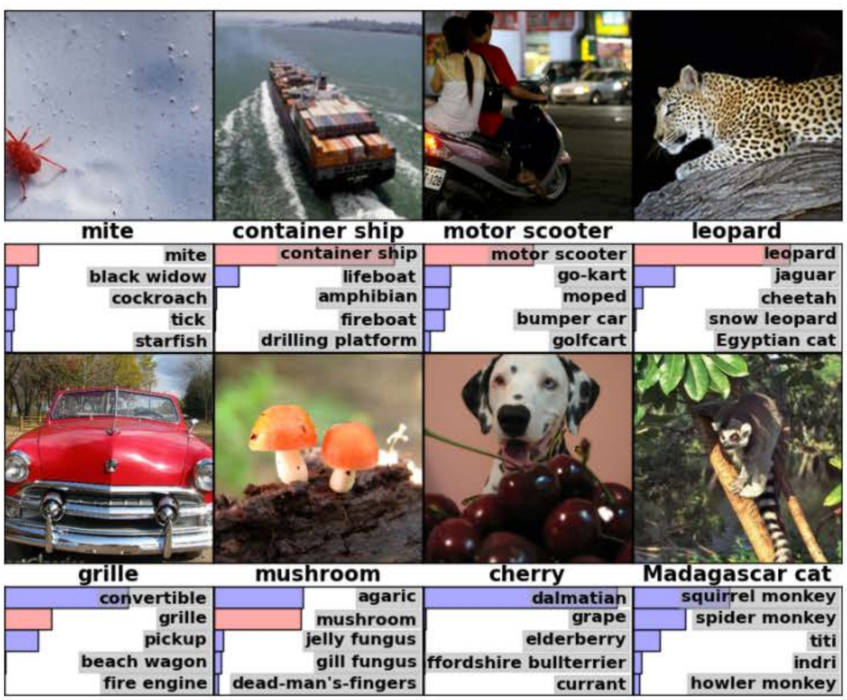
http://ganguli-gang.stanford.edu

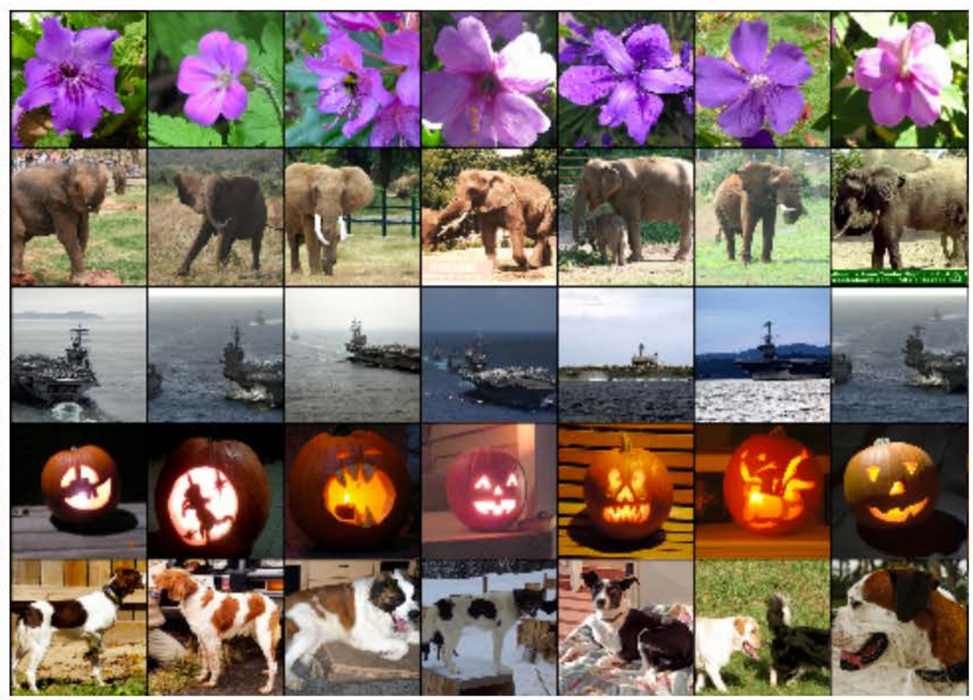
Twitter: @SuryaGanguli

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, 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.

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

Talk Outline

Generalization: How can networks learn probabilistic models of the world and imagine things they have not explicitly been taught?

Modelling arbitrary probability distributions using non-equilibrium thermodynamics,
J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, ICML 2015.

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.

Learning deep generative models by reversing diffusion

with Jascha Sohl-Dickstein
Eric Weiss, Niru Maheswaranathan

Goal: Model complex probability distributions – i.e. the distribution over natural images.

Once you have learned such a model, you can use it to:

Imagine new images
Modify images
Fix errors in corrupted images

Goal: achieve highly flexible but also tractable probabilistic generative models of data

- Physical motivation
  - Destroy structure in data through a diffusive process.
  - Carefully record the destruction.
  - Use deep networks to **reverse time and create structure from noise.**
- Inspired by recent results in non-equilibrium statistical mechanics which show that entropy can transiently decrease for short time scales (violations of second law)

Physical Intuition: Destruction of Structure through Diffusion

- Dye density represents probability density
- Goal: Learn structure of probability density
- Observation: Diffusion destroys structure

Data distribution → Uniform distribution

Physical Intuition: Recover Structure by Reversing Time

- What if we could reverse time?
- Recover data distribution by starting from uniform distribution and running dynamics backwards

Data distribution ← Uniform distribution

Physical Intuition: Recover Structure by Reversing Time

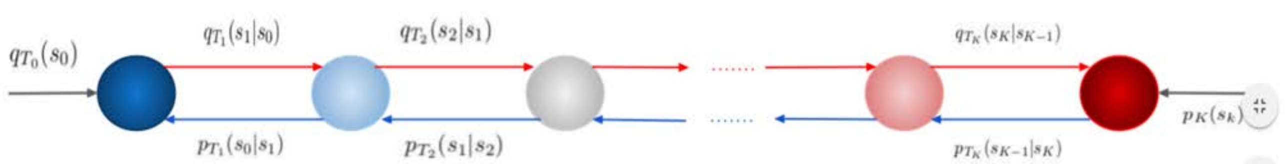
- What if we could reverse time?
- Recover data distribution by starting from uniform distribution and running dynamics backwards (using a trained deep network)

Data distribution ← Uniform distribution

Reversing time using a neural network

Finite time diffusion steps

Complex Data Distribution



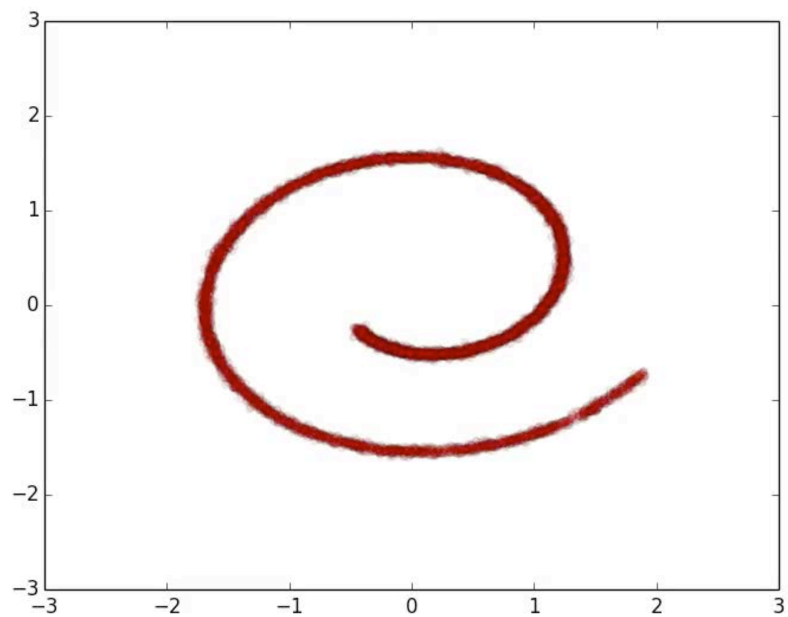
Simple Distribution

Neural network processing

Minimize the Kullback-Leibler divergence between forward and backward trajectories over the weights of the neural network

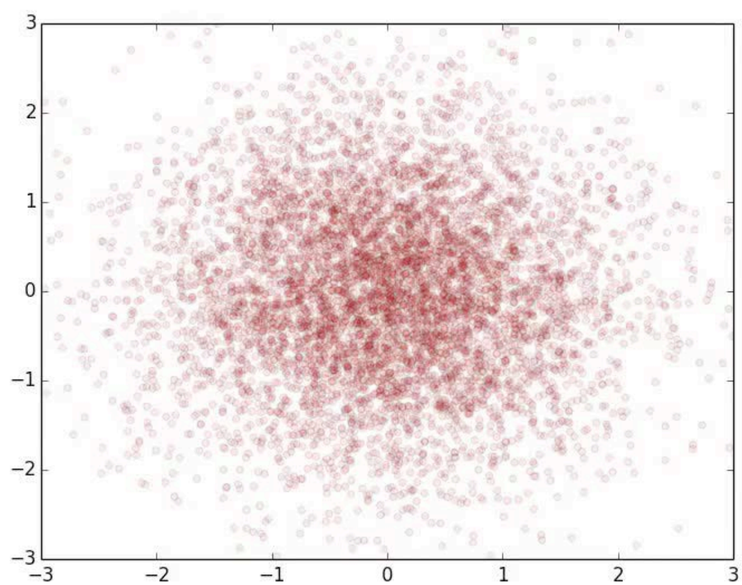
Swiss Roll

- Forward diffusion process
  - Start at data
  - Run Gaussian diffusion until samples become Gaussian blob



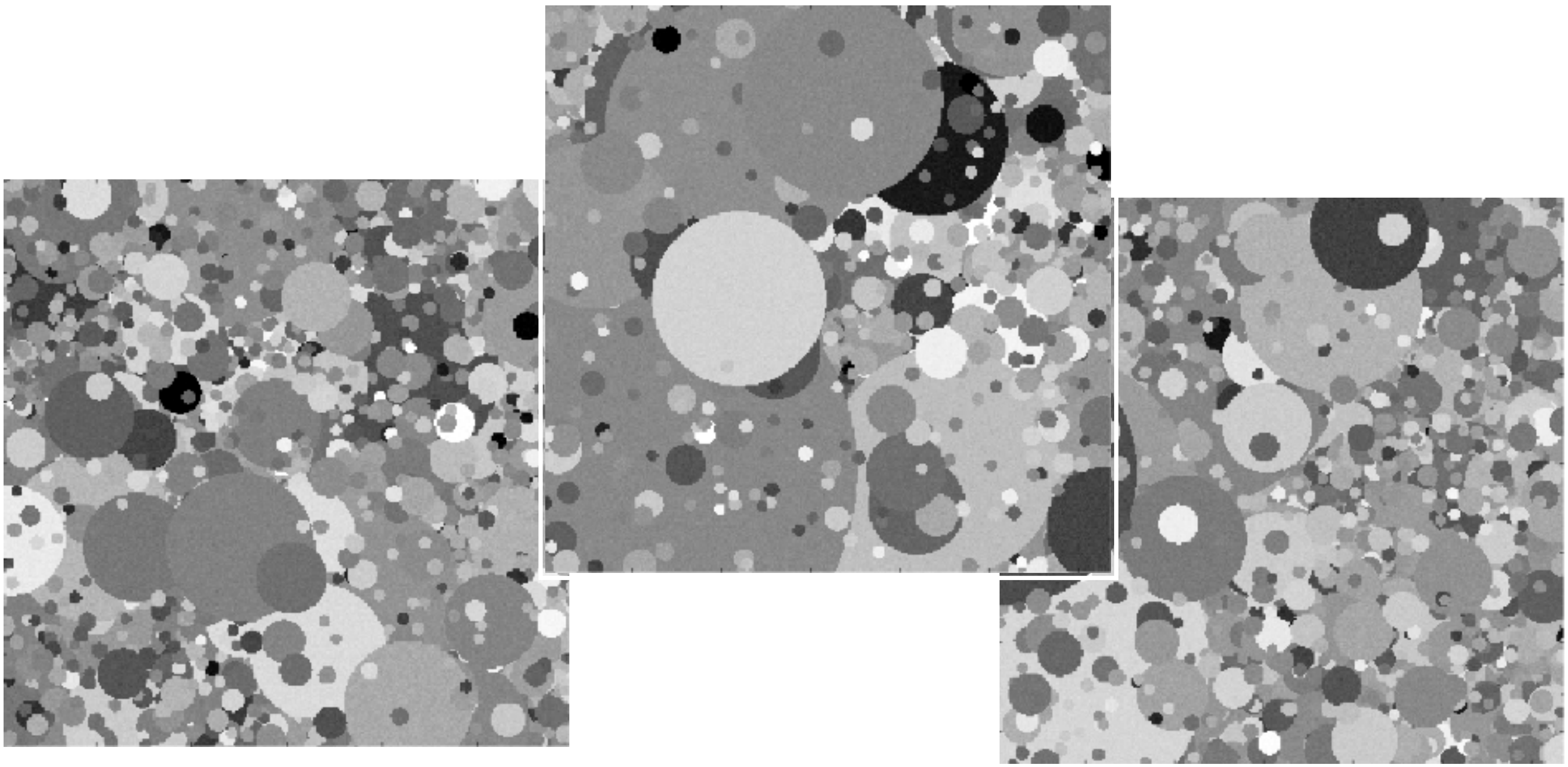
Swiss Roll

- Reverse diffusion process
  - Start at Gaussian blob
  - Run Gaussian diffusion until samples become data distribution



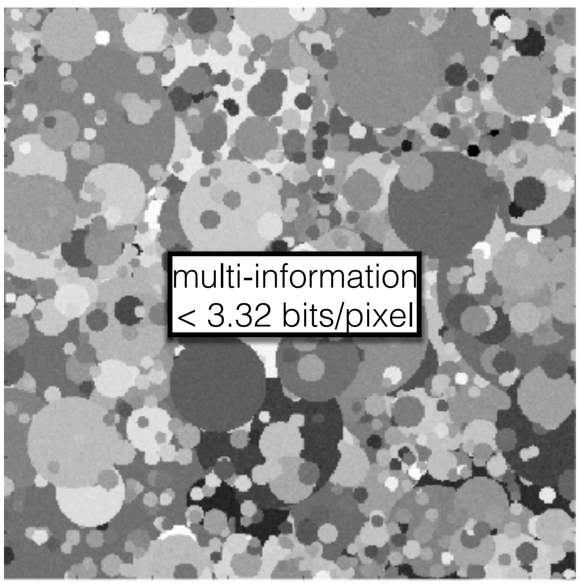
Dead Leaf Model

- Training data

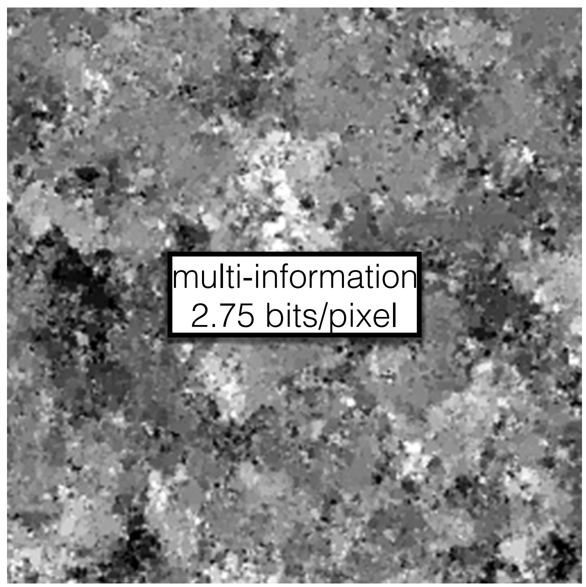


Dead Leaf Model

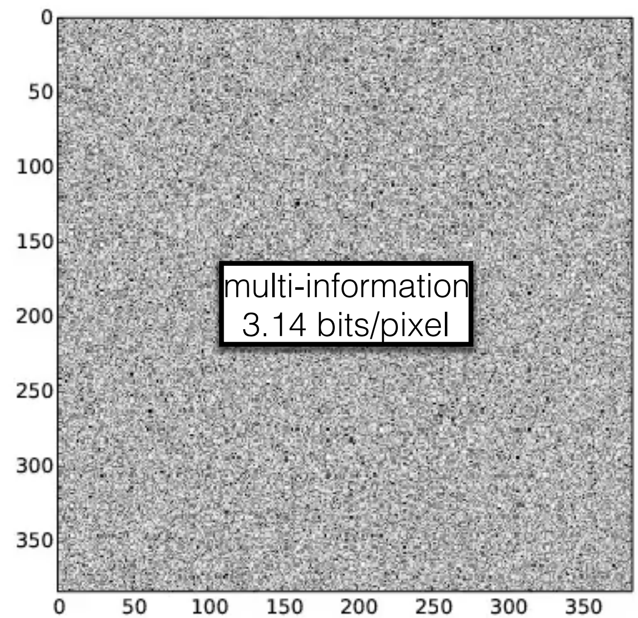
- Comparison to state of the art



Training Data



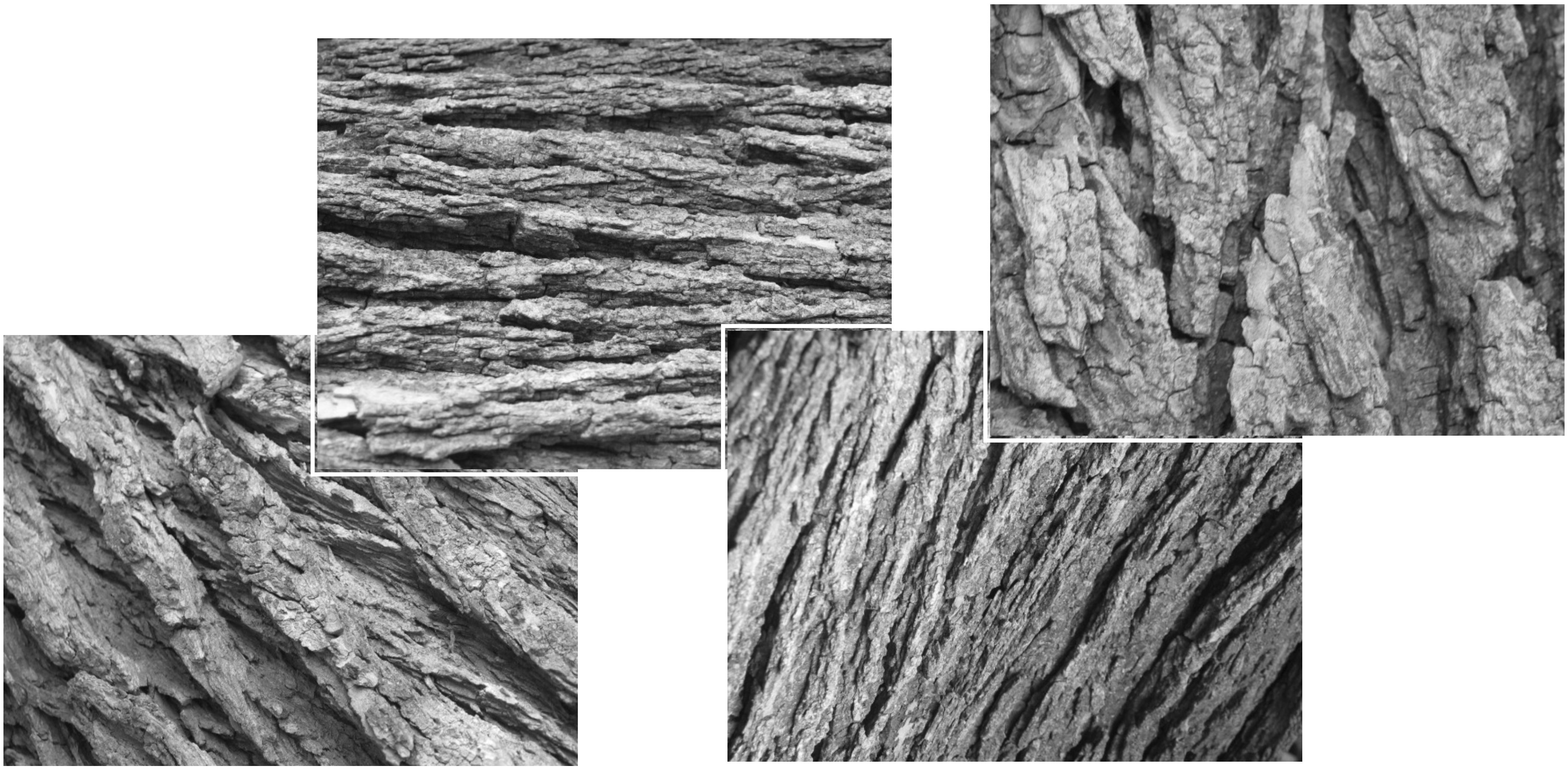
Sample from
[Theis *et al*, 2012]



Sample from
diffusion model

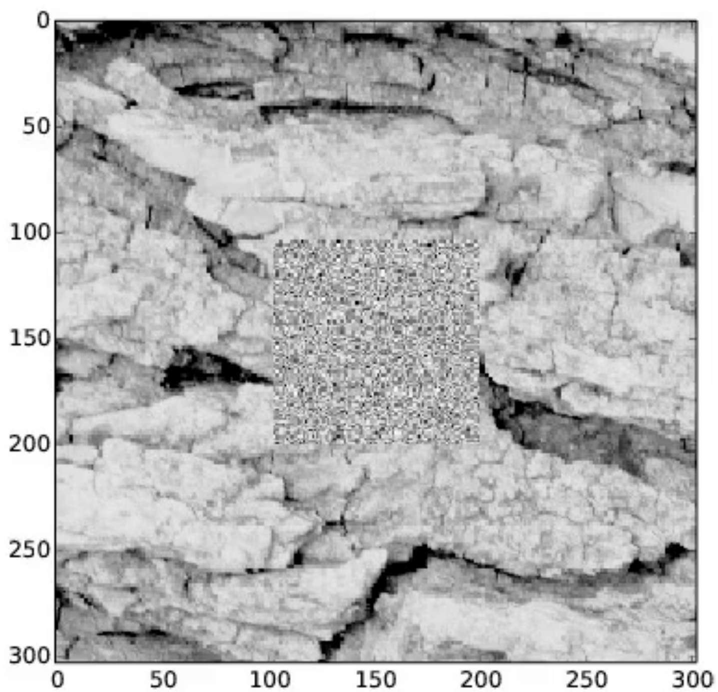
Natural Images

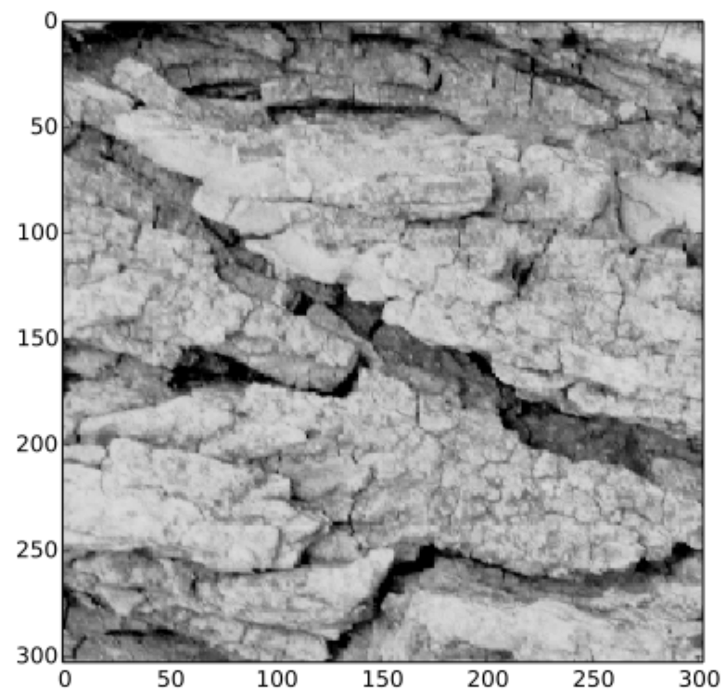
- Training data



Natural Images

- Inpainting





A key idea: solve the mixing problem during learning

- We want to model a complex multimodal distribution with energy barriers separating modes
- Often we model such distributions as the stationary distribution of a stochastic process
- But then mixing time can be long – exponential in barrier heights
- Here: Demand that we get to the stationary distribution in a finite time transient non-eq process!
- Build in this requirement into the learning process to obtain non-equilibrium models of data

Talk Outline

Generalization: How can networks learn probabilistic models of the world and imagine things they have not explicitly been taught?

Modelling arbitrary probability distributions using non-equilibrium thermodynamics,
J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, ICML 2015.

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.

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?

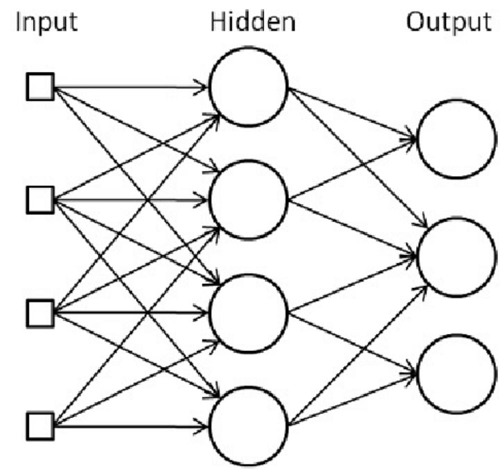
Exponential expressivity in deep neural networks through transient chaos, B. Poole, S. Lahiri,M. Raghu, J. Sohl-Dickstein, S. Ganguli, NIPS 2016.

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)

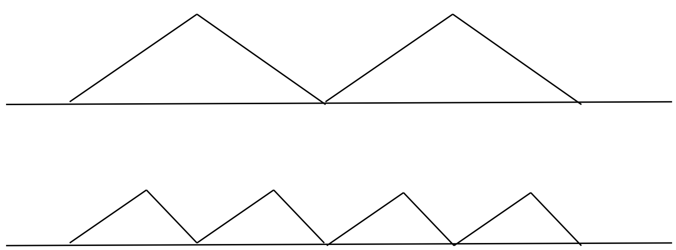
Intellectual traditions in boolean circuit theory: parity function is such a function for boolean circuits.

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.



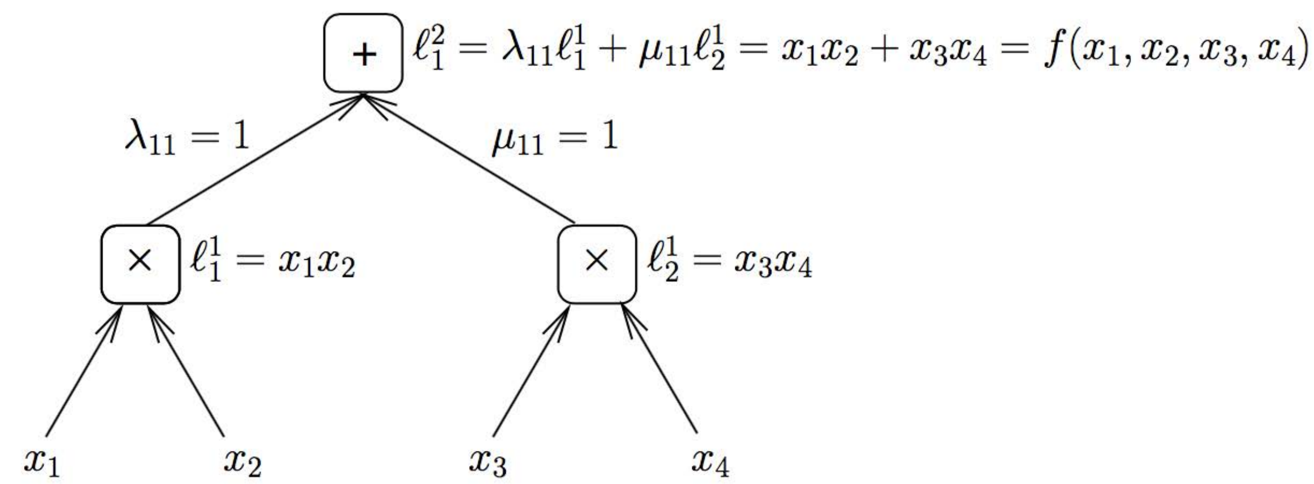
Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio.
On the number of linear regions of deep neural networks, NIPS 2014

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?

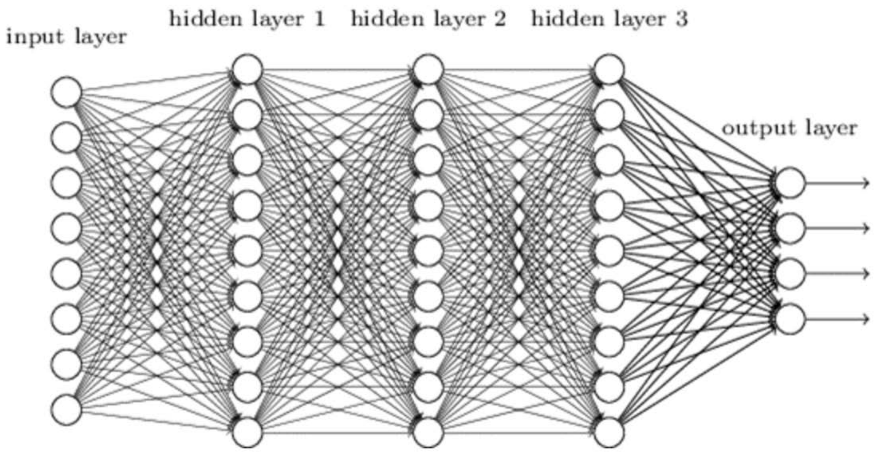
If so we would like a theory of deep neural expressivity that demonstrates this for

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



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

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

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

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

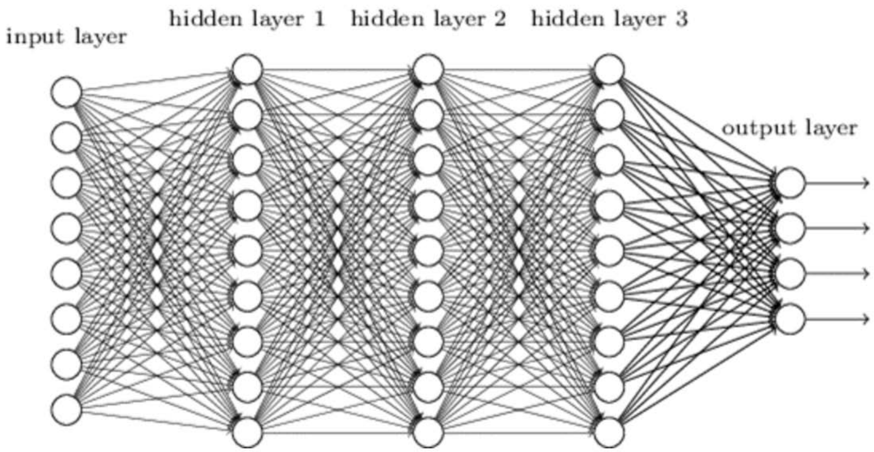
Structure:

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

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

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

Emergent, deterministic signal propagation in random neural networks



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

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

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

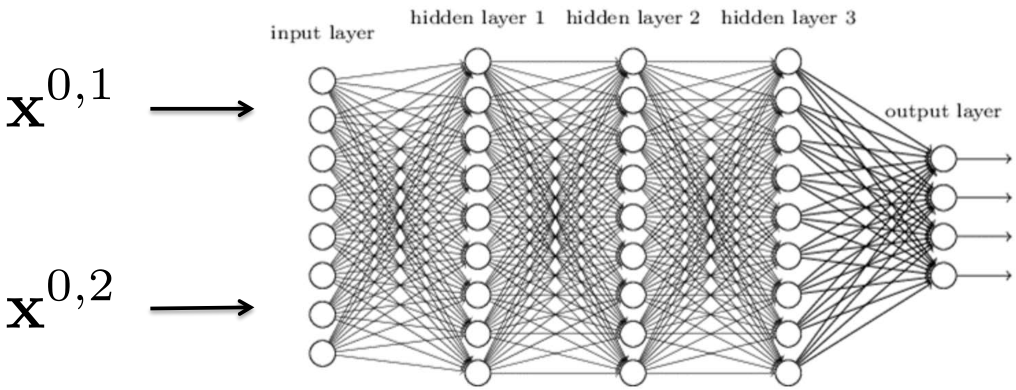
$$\mathbf{h}^l = \mathbf{W}^l \, \mathbf{x}^{l-1} + \mathbf{b}^l$$

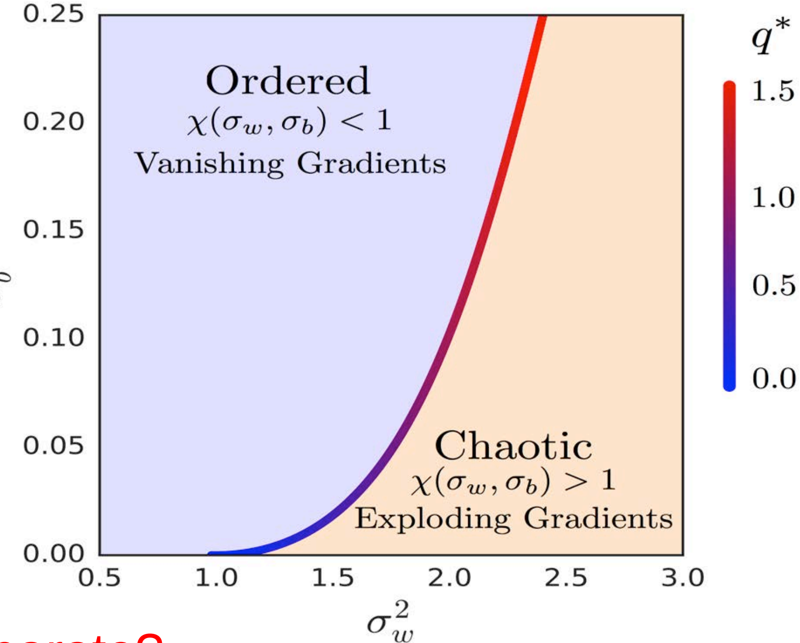
Question: how do simple input manifolds propagate through the layers?

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?

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

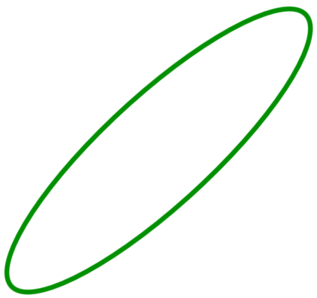
χ is the mean squared singular value of the Jacobian across 1 layer

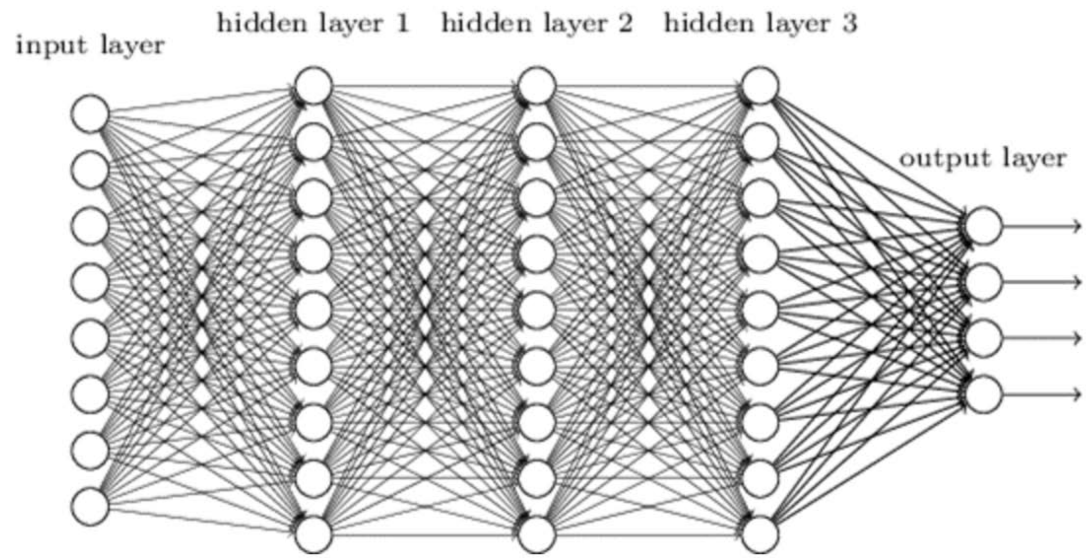
$\chi < 1$: nearby points come closer together; gradients exponentially vanish
$\chi > 1$: nearby points are driven apart; gradients exponentially explode

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

Propagation of a manifold through a deep network

$$\mathbf{x}^0(\theta)$$





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

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

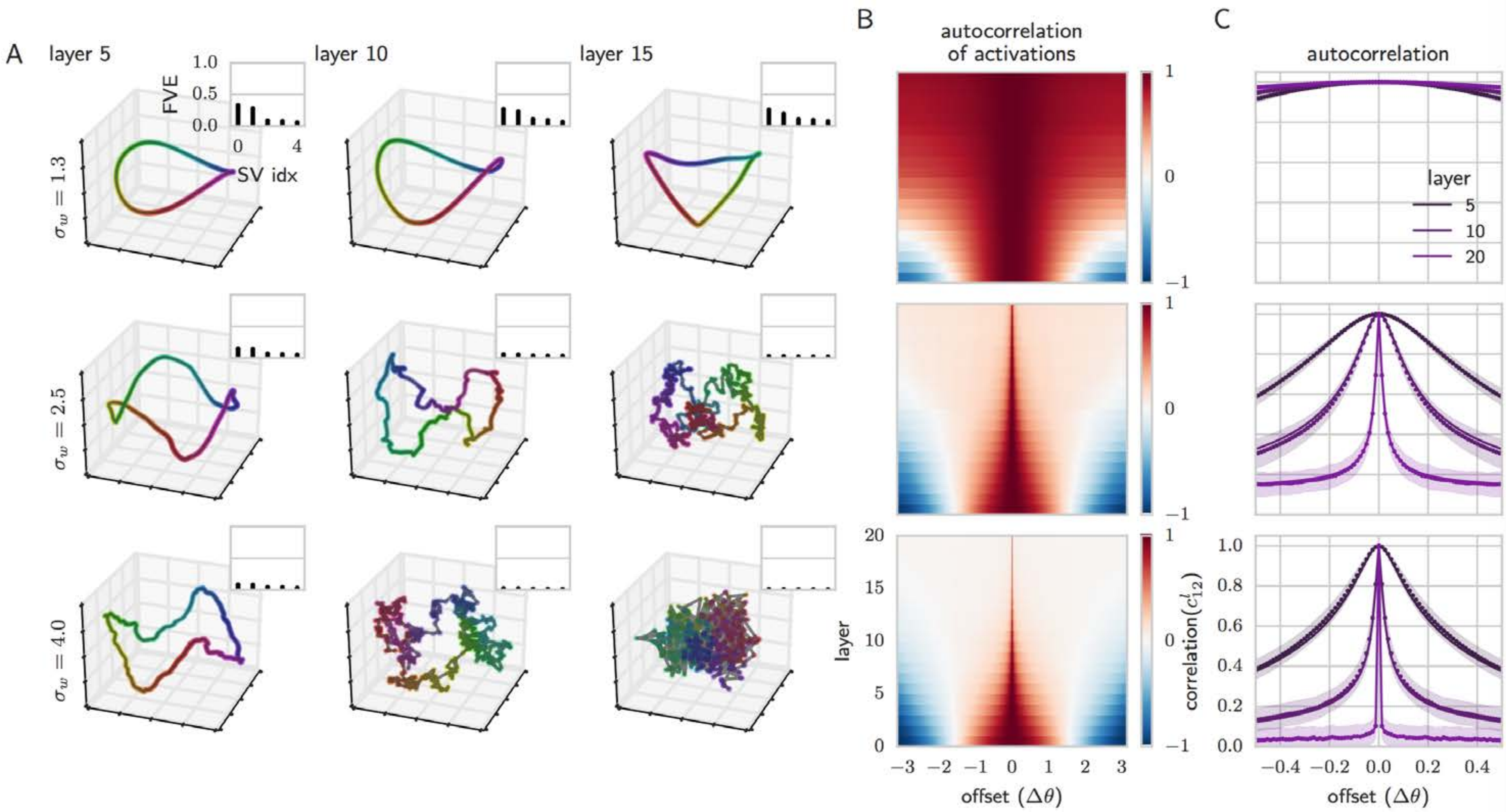
Or autocorrelation function:

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

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

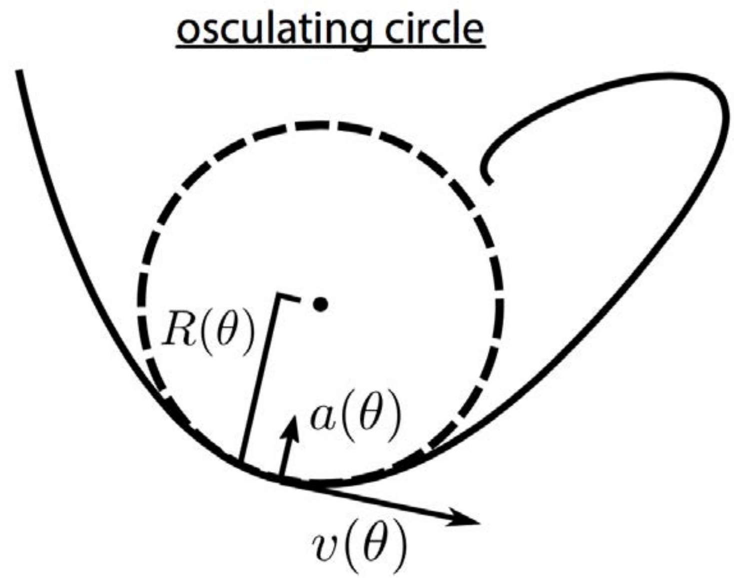


Riemannian geometry: Extrinsic Gaussian Curvature

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

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

$$\mathbf{a}(\theta) = \frac{\partial \mathbf{v}(\theta)}{\partial \theta}$$ 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 **h**(θ), with the same velocity and acceleration.

The extrinsic curvature κ(θ) is the inverse of the radius of this circle.

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

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

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

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

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

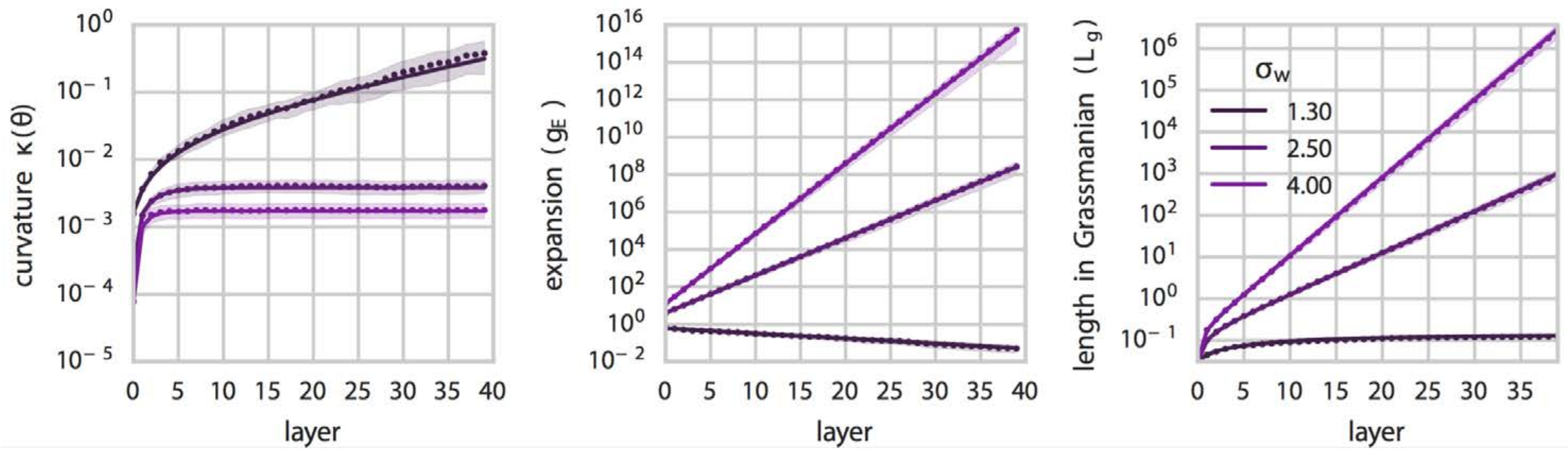
$$\chi_2 = \sigma_w^2 \int \mathcal{D}z \left[\phi''\left(\sqrt{q^*}z\right)\right]^2$$

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

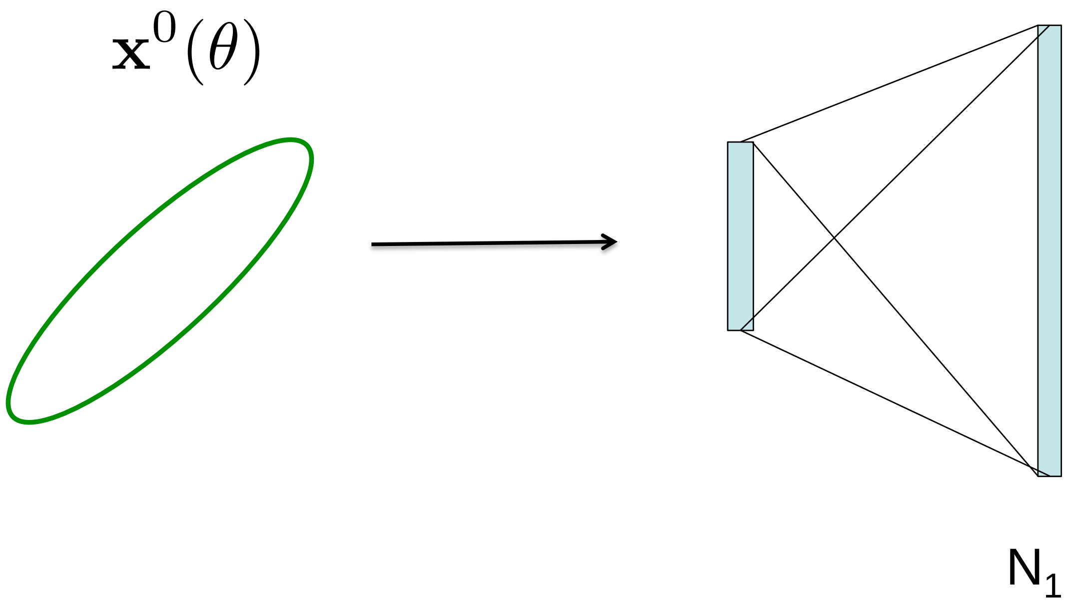


Unlike linear expansion, deep neural signal propagation can:

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!

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.

http://ganguli-gang.stanford.edu