Trends in ML/AI: Today's Frontiers and Challenges and Opportunities for the Future

Surya Ganguli

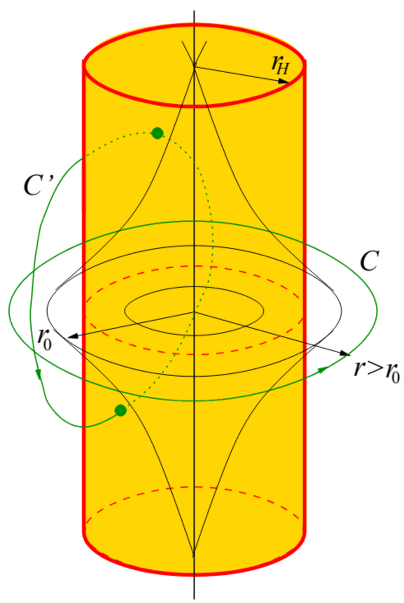
Stanford University
Depts. of Applied Physics,
Neurobiology,
Electrical Engineering, and
Computer Science

http://ganguli-gang.stanford.edu **Twitter: @SuryaGanguli**

My first paper in my former life as a string theorist

Holographic Protection of Chronology in Universes of the Godel Type

Edward Boyda, Surya Ganguli, Petr Horava, Uday Varadarajan (Berkeley)



Outline: today's frontiers and tomorrow challenges/opportunities

(1) Alternate physical implementations of ML algorithms: negotiating high dimensional error landscapes with atoms and photons in open dissipative open systems.

(2) The physical origins of generative AI: from diffusion models to societal implications.

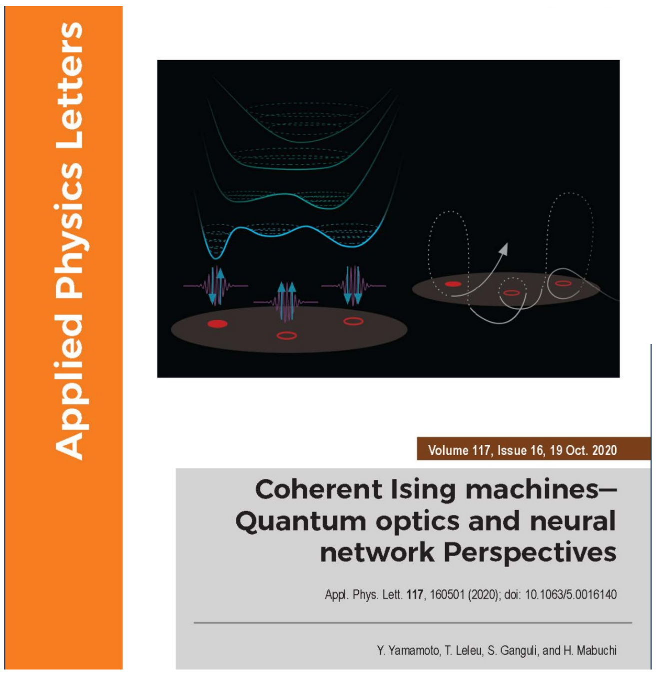
(3) Mechanistic interpretability: a unification of neuroscience, physics and AI with applications to alignment and government regulation

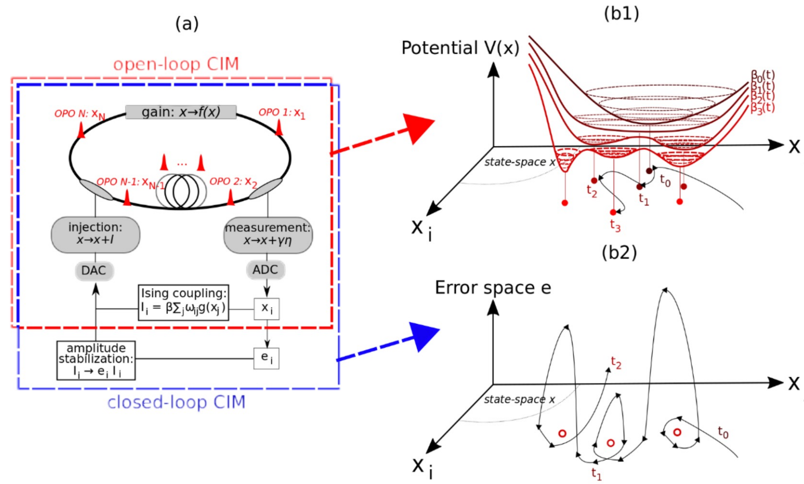
(4) Theory: the generalization puzzle: how can huge networks generalize without overfitting.

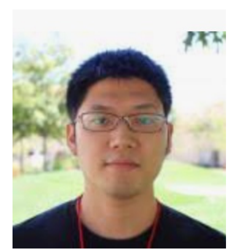
(5) Unsustainable neural scaling laws and the combatting the data hungriness of modern AI

(6) The mystery of structured sequences: the unreasonable effectiveness of predicting every next word on the internet (+ human feedback). LLMs as the next challenge/opportunity for theory and society.

Statistical mechanics of high dimensional optimization landscapes in the Coherent Ising Machine







Joint work with: Atushi Yamamura Hideo Mabuchi

Many Combinatorial optimization problems can encoded in an Ising formulation

$$E(x) = -\sum_{ij} J_{ij} x_i x_j \qquad x_i \in \{\pm 1\}$$

frontiers in PHYSICS

REVIEW ARTICLE
published: 12 February 2014
doi: 10.3389/fphy.2014.00005

Ising formulations of many NP problems

Andrew Lucas*

Lyman Laboratory of Physics, Department of Physics, Harvard University, Cambridge, MA, USA

Edited by:
Jacob Biamonte, ISI Foundation, Italy

Reviewed by:
Mauro Faccin, ISI Foundation, Italy
Ryan Babbush, Harvard University, USA
Bryan A. O'Gorman, NASA, USA

We provide Ising formulations for many NP-complete and NP-hard problems, including all of Karp's 21 NP-complete problems. This collects and extends mappings to the Ising model from partitioning, covering, and satisfiability. In each case, the required number of spins is at most cubic in the size of the problem. This work may be useful in designing adiabatic quantum optimization algorithms.

Keywords: spin glasses, complexity theory, adiabatic quantum computation, NP, algorithms

Examples:			
	Number partitioning	Satisfiability	Traveling salesman
	Graph partitioning	Matching	Steiner trees
	Clique finding	Set covering	Knapsack problems
	Set packing	Vertex cover	Graph coloring

Three energy functions of interest

$$E(x) = -\sum_{ij} J_{ij} x_i x_j \qquad x_i \in \{\pm 1\}$$

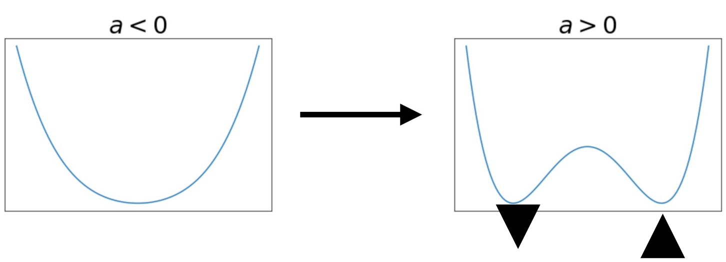
Ising energy
Encodes combinatorial problem of interest

$$E(x) = -\sum_{ij} J_{ij} x_i x_j \qquad x_i \in \mathbb{R} \quad x^T x = 1$$

Spectral relaxation; solution: *maximal* eigenvector of J

$$E(x) = \frac{1}{4}\sum_i x_i^4 - \frac{a}{2}\sum_i x_i^2 - \sum_{ij} J_{ij} x_i x_j \qquad x_i \in \mathbb{R}$$

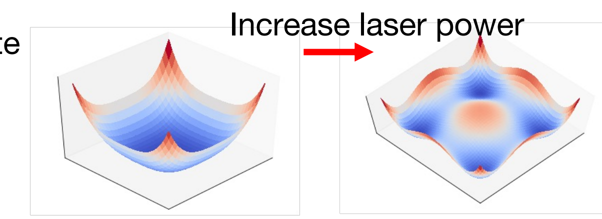
Classical version of CIM energy
a = laser pump power parameter
x = x quadrature of an OPO



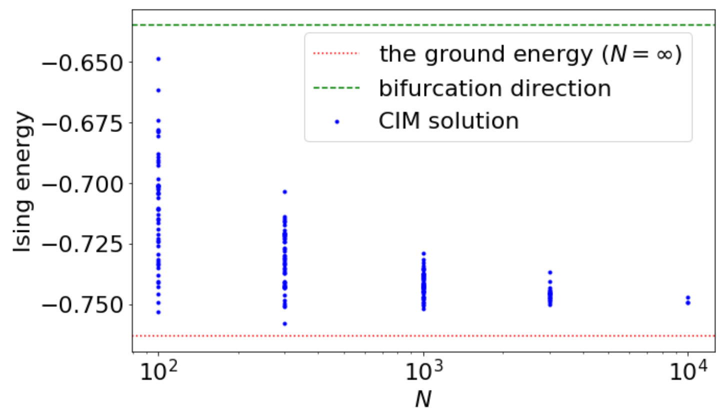
Approaching the SK spin glass solution through adiabatic evolution of the CIM

Sherrington Kirkpatrick (SK) spin glass: J_{ij} drawn as a Gaussian: $\mathcal{N}(0, 1/N)$

A result: adiabatic evolution of CIM to find SK low energy state



1) Start from the origin at a small value of laser pump power
2) Increase the pump power a small amount
3) Minimize energy via gradient descent
4) Repeat

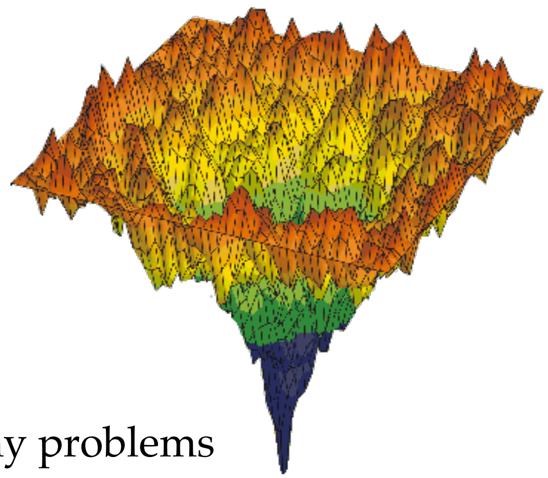


CIM adiabatic evolution outperforms the spectral solution and comes close to the theoretically predicted SK ground state energy

What makes this possible?
What is the shape of the energy landscape?
How does it change with laser pump power?
How can exploit our understanding of this changing geometry to determine optimal annealing schedule?

High dimensional nonconvex optimization

It is often thought that local minima at high error stand as as a major impediment to non-convex optimization.



In random non-convex error surfaces over high dimensional spaces, local minima at high error are exponentially rare in the dimensionality.

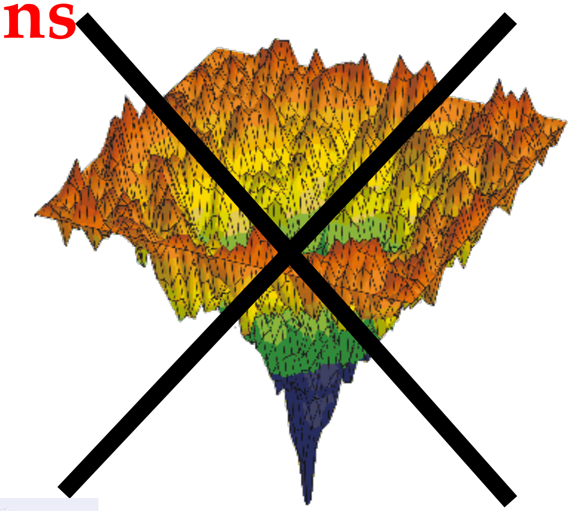
Instead saddle points proliferate.

We demonstrated this picture indeed occurs in many problems of relevance to deep learning and artificial intelligence,

And we developed an algorithm that rapidly escapes saddle points in high dimensional spaces.

Identifying and attacking the saddle point problem in high dimensional non-convex optimization.
Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, Yoshua Bengio. NIPS 2014

General properties of error landscapes in high dimensions



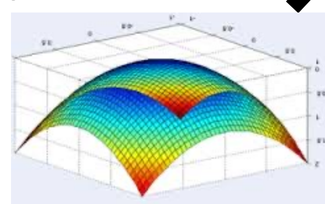
From statistical physics:

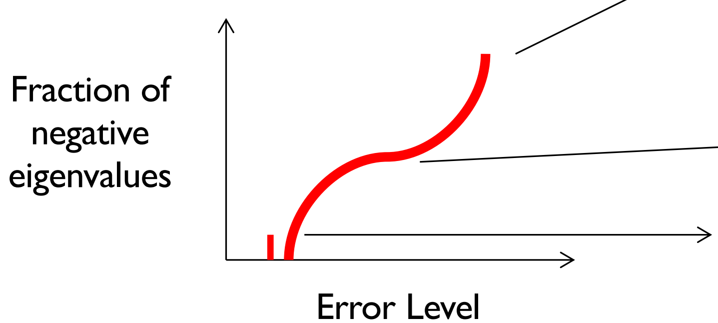
Consider a random Gaussian error landscape over N variables.

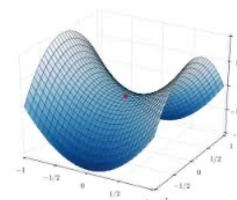
Let x be a critical point.

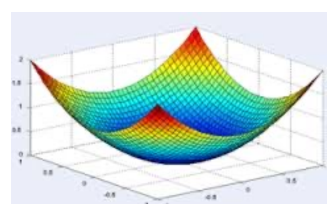
Let E be its error level.

Let f be the fraction of negative curvature directions.



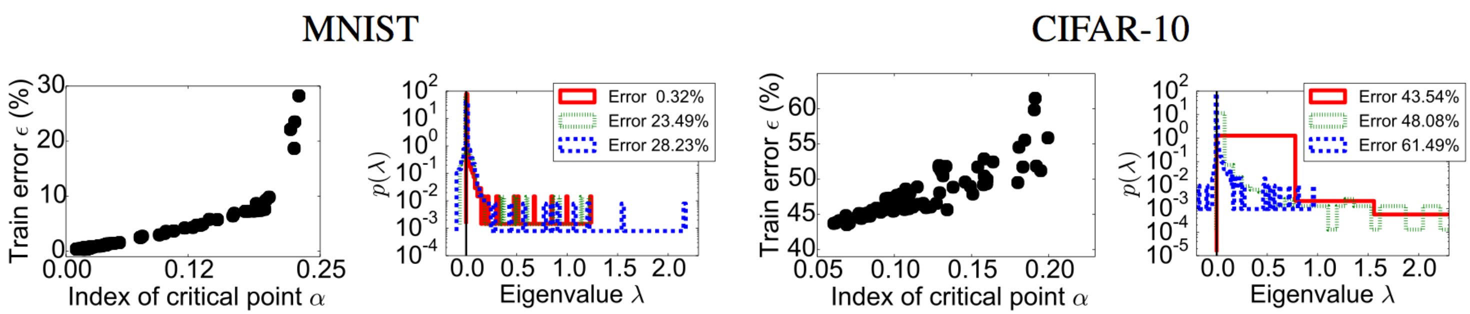






Bray and Dean, Physical Review Letters, 2007

Properties of Error Landscapes on the Synaptic Weight Space of a Deep Neural Net



Qualitatively consistent with the
statistical physics theory of random error landscapes

Identifying and attacking the saddle point problem in high dimensional non-convex optimization.
Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, Yoshua Bengio. NIPS 2014

Understanding the changing energy landscape of the CIM under adiabatic evolution of the pump parameter

$$E(x) = \frac{1}{4}\sum_i x_i^4 - \frac{a}{2}\sum_i x_i^2 - \sum_{ij} J_{ij} x_i x_j \qquad x_i \in \mathbb{R}$$

Sherrington Kirkpatrick (SK) spin glass: J_{ij} drawn as a Gaussian: $\mathcal{N}(0, 1/N)$

Questions: In terms of:

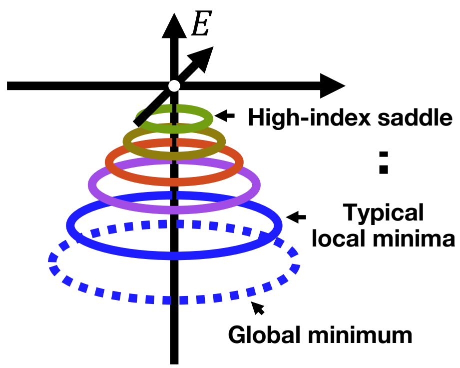
1) CIM energy E; 2) radial distance from origin; 3) pump power a:

Where does the global minimum lie?

Where do the most likely local minima lie?

Where do the most likely saddle points of a given index lie?

Where do the lowest energy saddle points of a given index lie?

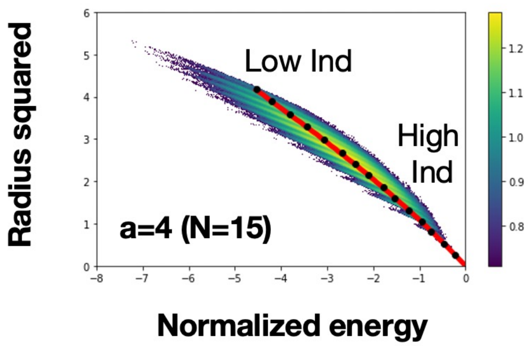


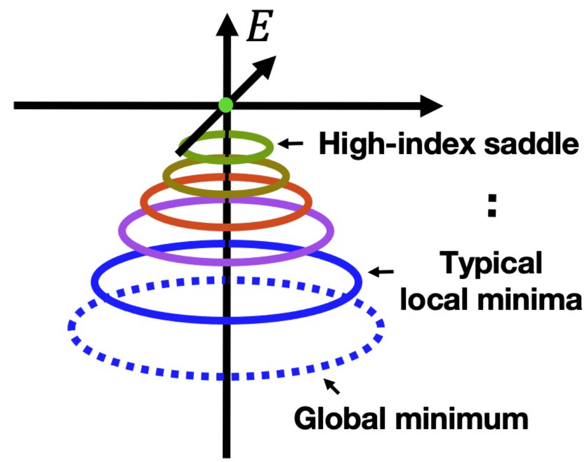
Given a critical point (minimum or saddle) what is the distribution of OPO quadrature x_i?

What is the eigenvalue spectrum of the Hessian of a critical point
as a function of its index, energy and radius?

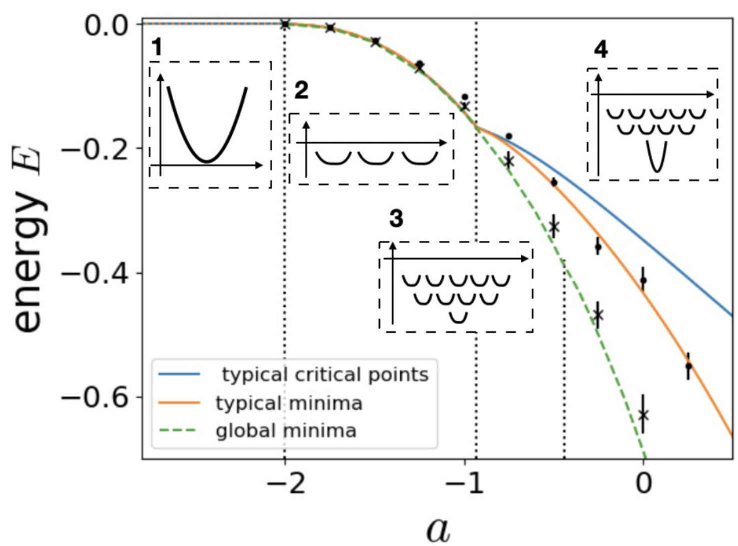
Geometry of the changing CIM Energy landscape as a function of laser pump power a

$$E(x) = \frac{1}{4}\sum_i x_i^4 - \frac{a}{2}\sum_i x_i^2 - \sum_{ij} J_{ij} x_i x_j \qquad x_i \in \mathbb{R}$$



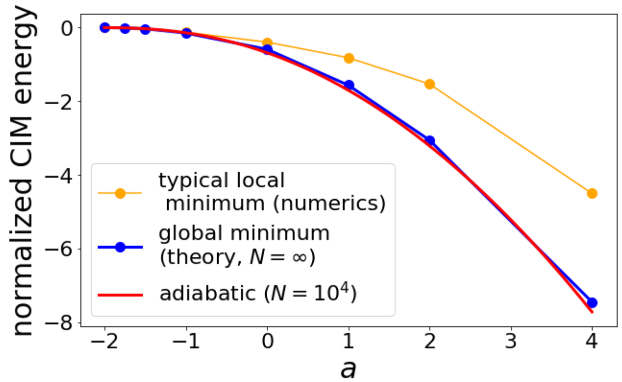


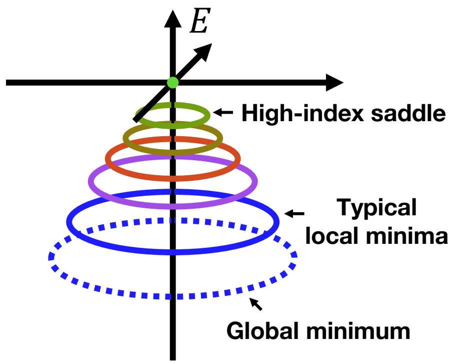
Phase transitions in the geometry of the CIM landscape



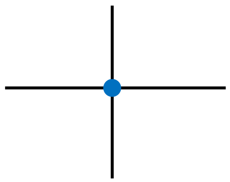
$a < -2$	$-2 < a < -0.93$	$-0.93 < a < -0.45$	$-0.45 < a$
Replica symmetric phase:	Replica symmetric breaking:	SUSY breaking:	Freezing:
A single stable local minimum	Exponentially many local minima All of similar energy densities All are marginally stable “Flat landscape”	Exponentially many local minima Most local minima marginally stable Global minimum marginally stable Range of energies Global min < Typical local min	Global minimum is no Longer marginally stable And is fully stable.

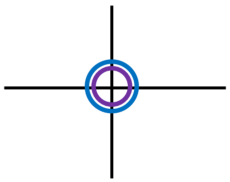
Adiabatic evolution tracks the global minimum of CIM energy

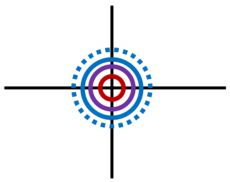


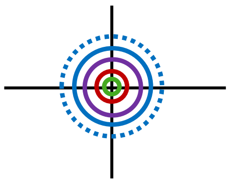


$a < -2$	$-2 < a < -0.93$	$-0.93 < a < -0.45$	$-0.45 < a$
Replica symmetric phase:	Replica symmetric breaking:	SUSY breaking:	Freezing:
A single stable local minimum	Exponentially many local minima All of similar energy densities All are marginally stable “Flat landscape”	Exponentially many local minima Most local minima marginally stable Global minimum marginally stable Range of energies Global min < Typical local min	Global minimum is no Longer marginally stable And is fully stable.







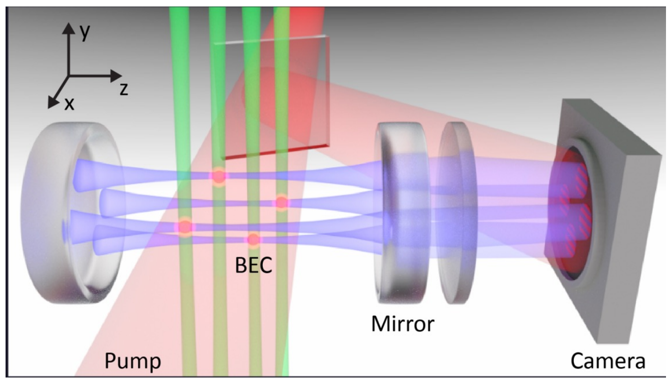


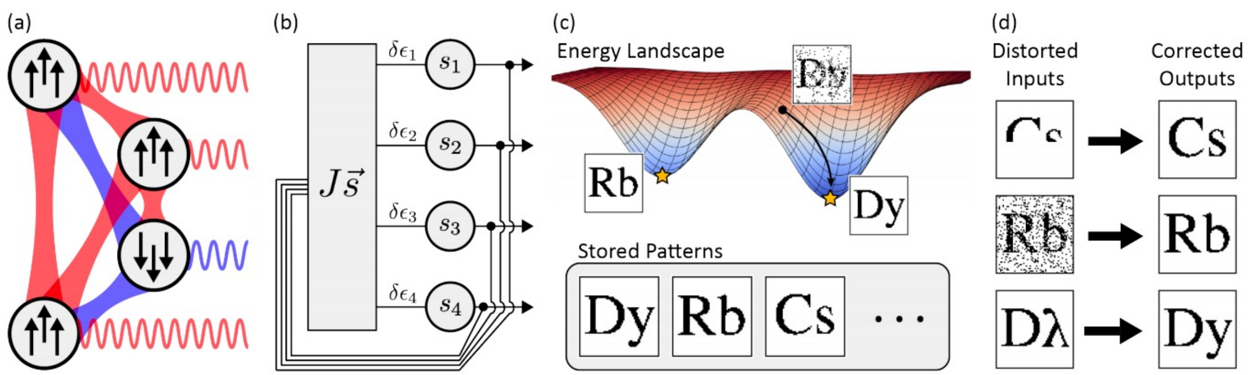
Enhancing Associative Memory Recall and Storage Capacity Using Confocal Cavity QED

Brendan P. Marsh, Yudan Guo, Ronen M. Kroeze, Sarang Gopalakrishnan, Surya Ganguli, Jonathan Keeling, and Benjamin L. Lev

Phys. Rev. X **11**, 021048 – Published 2 June 2021

Physics See synopsis: A Computer Memory Based on Cold Atoms and Light

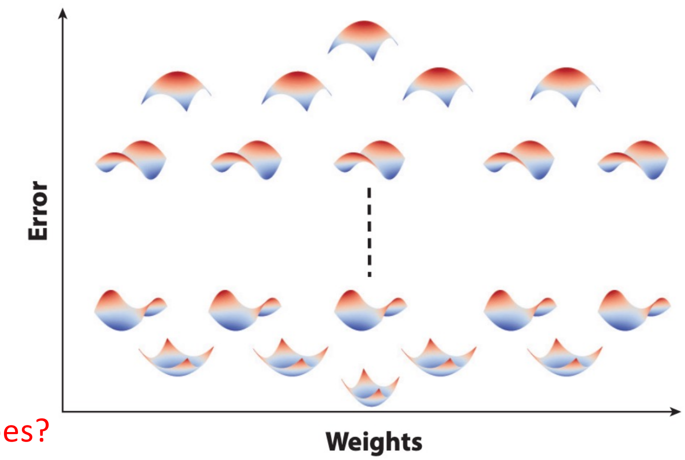




Entanglement and replica symmetry breaking in a driven-dissipative quantum spin glass

Arxiv:2307.1017

Brendan P. Marsh,[1,2] Ronen M. Kroeze,[3,2] Surya Ganguli,[1]
Sarang Gopalakrishnan,[4] Jonathan Keeling,[5] and Benjamin L. Lev[1,2,3]
[1]*Department of Applied Physics, Stanford University, Stanford CA 94305, USA*
[2]*E. L. Ginzton Laboratory, Stanford University, Stanford, CA 94305, USA*
[3]*Department of Physics, Stanford University, Stanford CA 94305, USA*
[4]*Department of Electrical and Computer Engineering, Princeton University, Princeton NJ 08544, USA*
[5]*SUPA, School of Physics and Astronomy, University of St. Andrews, St. Andrews KY16 9SS, United Kingdom*
(Dated: July 20, 2023)



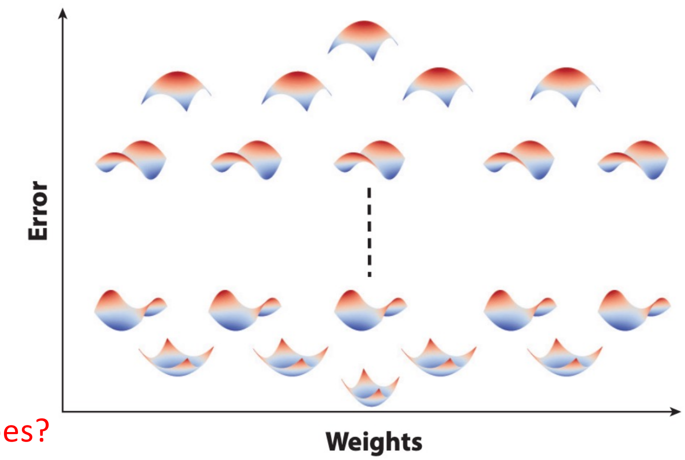
How does open dissipative quantum dynamics descend high dimensional error landscapes?

Entanglement and replica symmetry breaking in a driven-dissipative quantum spin glass

Arxiv:2307.1017

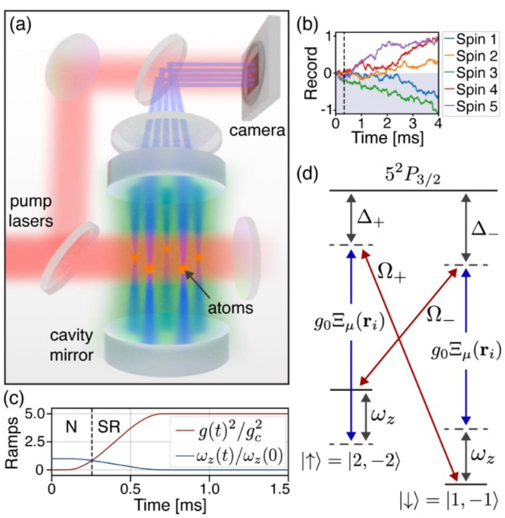
Brendan P. Marsh,[1,2] Ronen M. Kroeze,[3,2] Surya Ganguli,[1]
Sarang Gopalakrishnan,[4] Jonathan Keeling,[5] and Benjamin L. Lev[1,2,3]
[1]Department of Applied Physics, Stanford University, Stanford CA 94305, USA
[2]E. L. Ginzton Laboratory, Stanford University, Stanford, CA 94305, USA
[3]Department of Physics, Stanford University, Stanford CA 94305, USA
[4]Department of Electrical and Computer Engineering, Princeton University, Princeton NJ 08544, USA
[5]SUPA, School of Physics and Astronomy, University of St. Andrews, St. Andrews KY16 9SS, United Kingdom

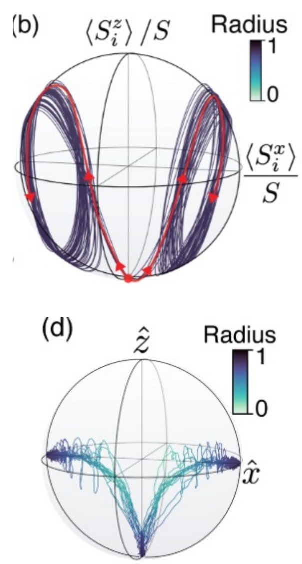
(Dated: July 20, 2023)

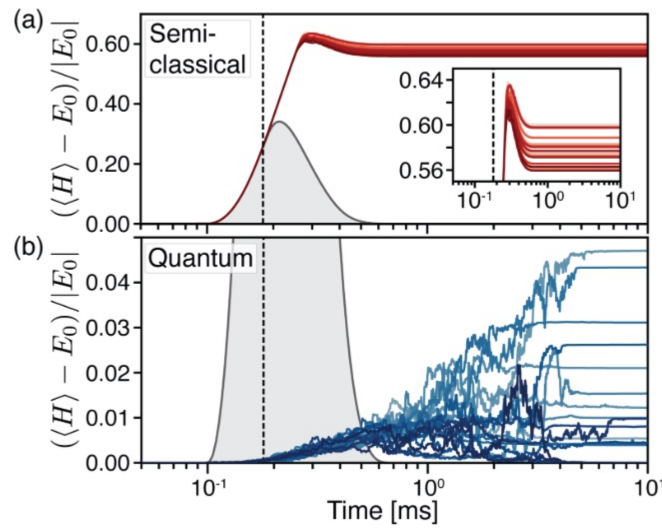


How does open dissipative quantum dynamics descend high dimensional error landscapes?

Quantum entanglement between multiple spins allows individual quantum trajectories to evade semi-classical energy barriers to achieve more optimal low energy states.



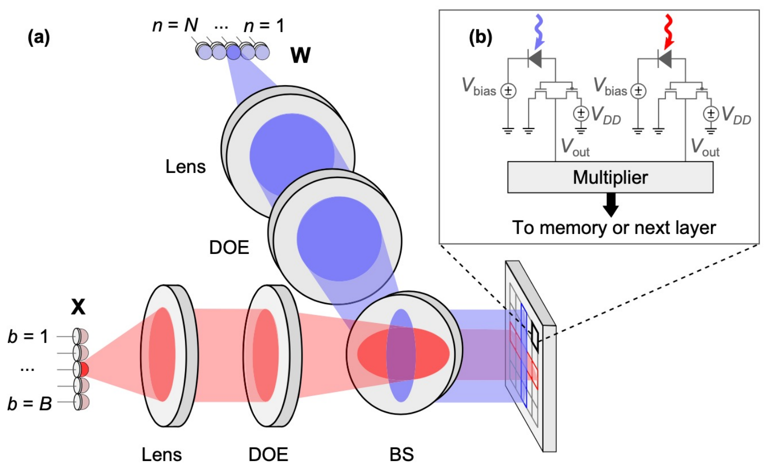


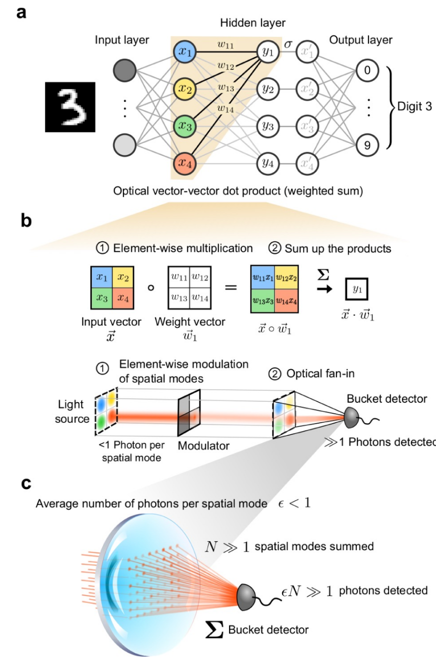


Fast high energy efficiency matrix multiplies with photons

Freely scalable and reconfigurable optical hardware for deep learning

Liane Bernstein[1,5]✉, Alexander Sludds[1,5]✉, Ryan Hamerly[1,2], Vivienne Sze[1], Joel Emer[3,4] & Dirk Englund[1]✉





An optical neural network using less than 1 photon p multiplication

Tianyu Wang ✉, Shi-Yuan Ma, Logan G. Wright, Tatsuhiro Onodera, Brian C. Richard & Peter L. McMahon ✉

Nature Communications **13**, Article number: 123 (2022) | Cite this article

Outline: today's frontiers and tomorrow challenges/opportunities

(1) Alternate physical implementations of ML algorithms: negotiating high dimensional error landscapes with atoms and photons in open dissipative open systems.

(2) The physical origins of generative AI: from diffusion models to societal implications.

(3) Mechanistic interpretability: a unification of neuroscience, physics and AI with applications to alignment and government regulation

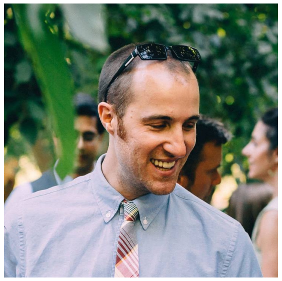
(4) Theory: the generalization puzzle: how can huge networks generalize without overfitting.

(5) Unsustainable neural scaling laws and the combatting the data hungriness of modern AI

(6) The mystery of structured sequences: the unreasonable effectiveness of predicting every next word on the internet (+ human feedback). LLMs as the next challenge/opportunity for theory and society.

The physical origins of generative AI: From nonequilbrium thermodynamics to artificial imagination

with Jascha Sohl-Dickstein
Eric Weiss, Niru Maheswaranathan



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

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

Physical Intuition: Recover Structure by Reversing Time

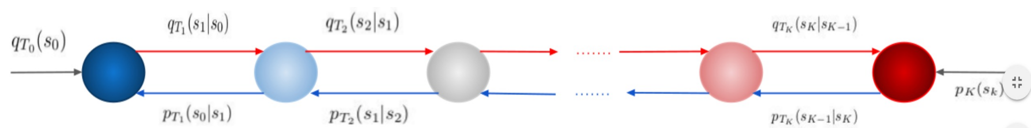
- What if we could reverse time?
- Recover data distribution by starting from uniform distribution and running dynamics backwards (using a trained deep network)

Data distribution ← Uniform distribution

Reversing time using a neural network

Finite time diffusion steps

Complex Data Distribution



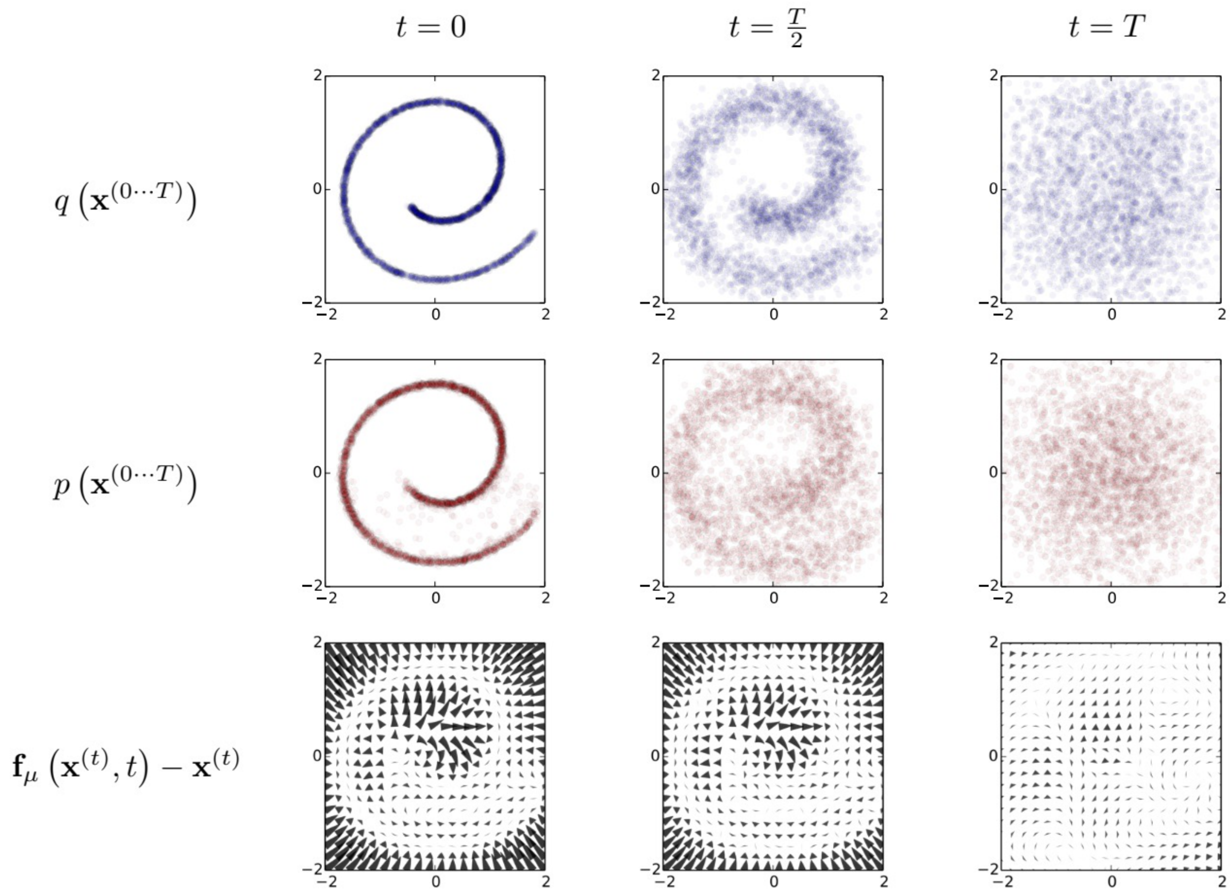
Simple Distribution

Neural network processing

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

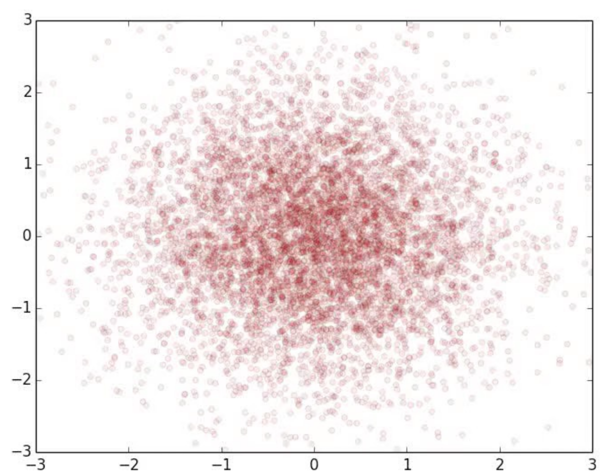
Swiss Roll

Deep Unsupervised Learning using Nonequilibrium Thermodynamics



Swiss Roll

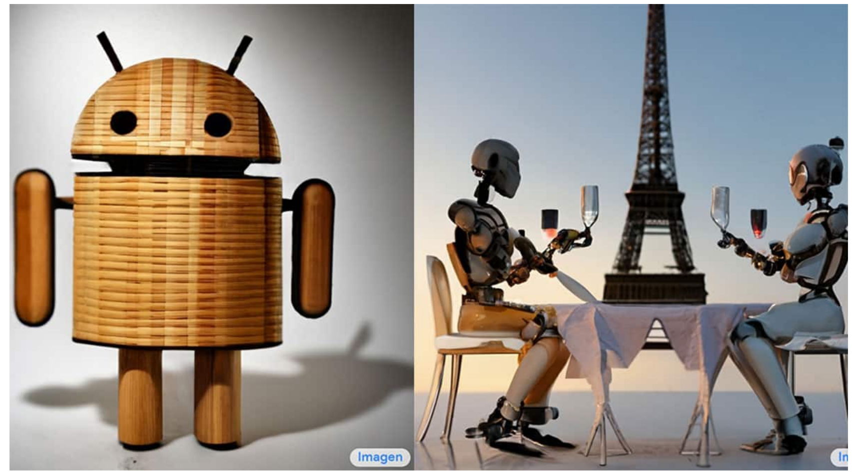
- Reverse diffusion process
  - Start at Gaussian blob
  - Run Gaussian diffusion until samples become data distribution

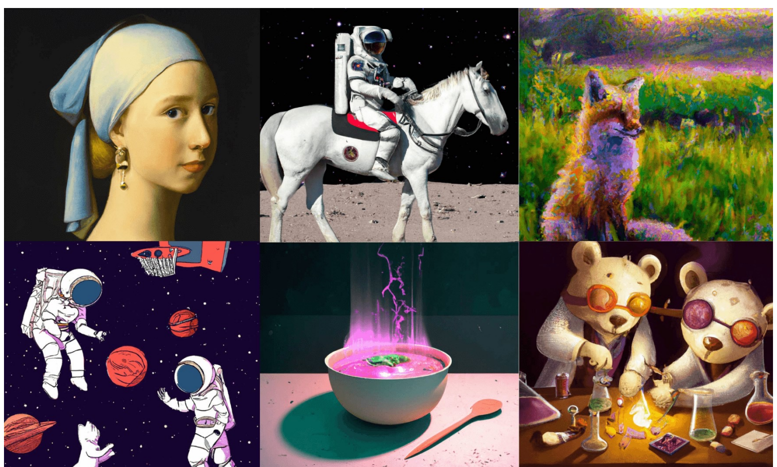


A key idea: solve the mixing problem during learning

- 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

Dalle2, MidJourney, ImageGen, Stable Diffusion





Societal implications:

Empowering human creativity

Artist’s rights / copyright law

Deep fakes

Outline: today's frontiers and tomorrow challenges/opportunities

(1) Alternate physical implementations of ML algorithms: negotiating high dimensional error landscapes with atoms and photons in open dissipative open systems.

(2) The physical origins of generative AI: from diffusion models to societal implications.

(3) Mechanistic interpretability: a unification of neuroscience, physics and AI with applications to alignment and government regulation

(4) Theory: the generalization puzzle: how can huge networks generalize without overfitting.

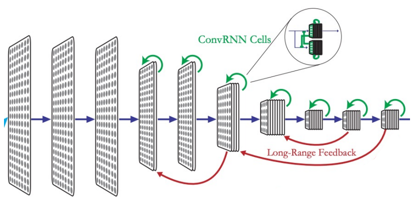
(5) Unsustainable neural scaling laws and the combatting the data hungriness of modern AI

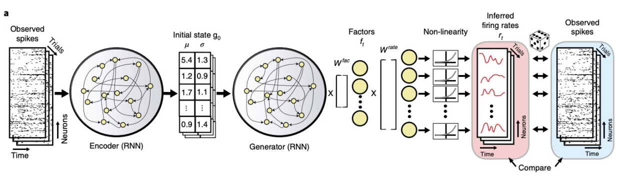
(6) The mystery of structured sequences: the unreasonable effectiveness of predicting every next word on the internet (+ human feedback). LLMs as the next challenge/opportunity for theory and society.

Mechanistic interpretability:
unifying neuroscience physics and AI

Macaque
10^9 neurons

Human
10^{11} neurons





Mechanistic Interpretability in Neuroscience:

Deep learning allows us to create highly accurate,
Predictive models of neural circuits.

But are we just replacing something we don't understand
(i.e. the brain) with something else we don't understand (our model of it)

Mechanistic Interpretability in AI:

How do we align what AI systems do with human values?

Well, what heck are AI systems actually doing?
How do they work?

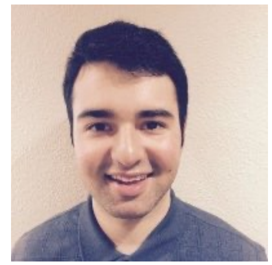
What do they actually understand?

Government regulation: auditing / editing AI systems.
Right to be forgotten.

Interpretable deep neural network models of the retinal response to natural movies

Lane McIntosh

Niru Maheswaranathan



Aran Nayebi

Hidenori Tanaka

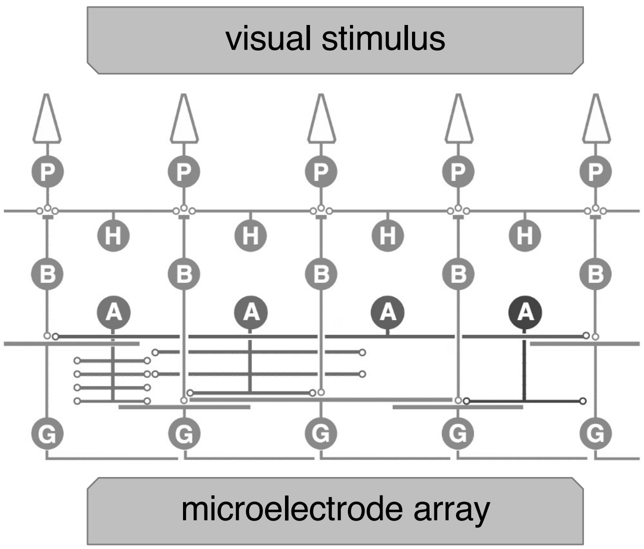
McIntosh, L.*, Maheswaranathan, N.*, Nayebi, A., Ganguli, S., Baccus, S.A. Deep Learning Models of the Retinal Response to Natural Scenes. NIPS 2016.

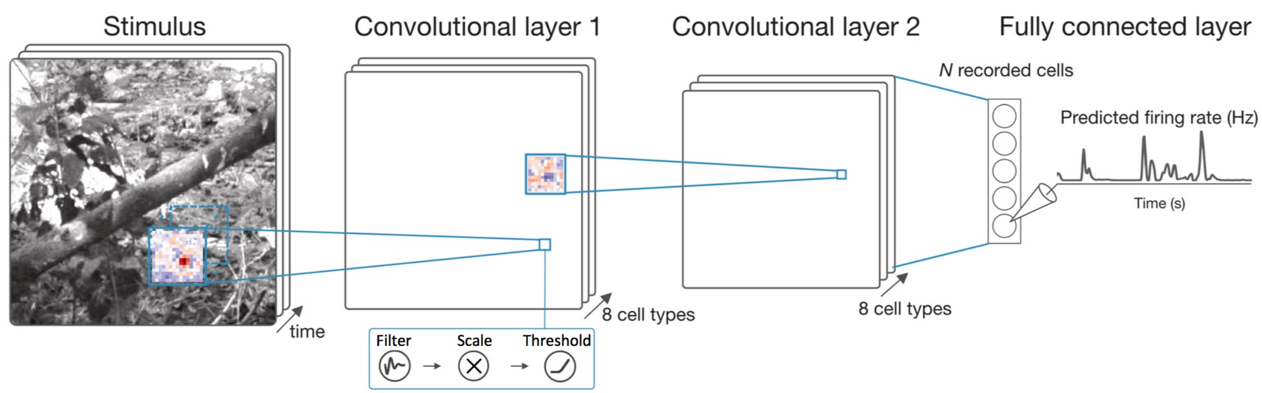
Maheswaranathan, N, Baccus, S. and Ganguli, S., Inferring hidden structure in multilayered neural circuits, PLOS Computational Biology, 2018.

From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction, Hidenori Tanaka, Aran Nayebi, Stephen A. Baccus, Surya Ganguli, NeurIPS 2018

Niru Maheswaranathan*, Lane McIntosh*, David B. Kastner, Josh Melander, Luke Brezovec, Aran Nayebi, Julia Wang, Surya Ganguli and Stephen A. Baccus, Interpreting the retinal neural code for natural scenes: From computations to neurons, Neuron 2023.

The retina: biological complexity versus model simplicity

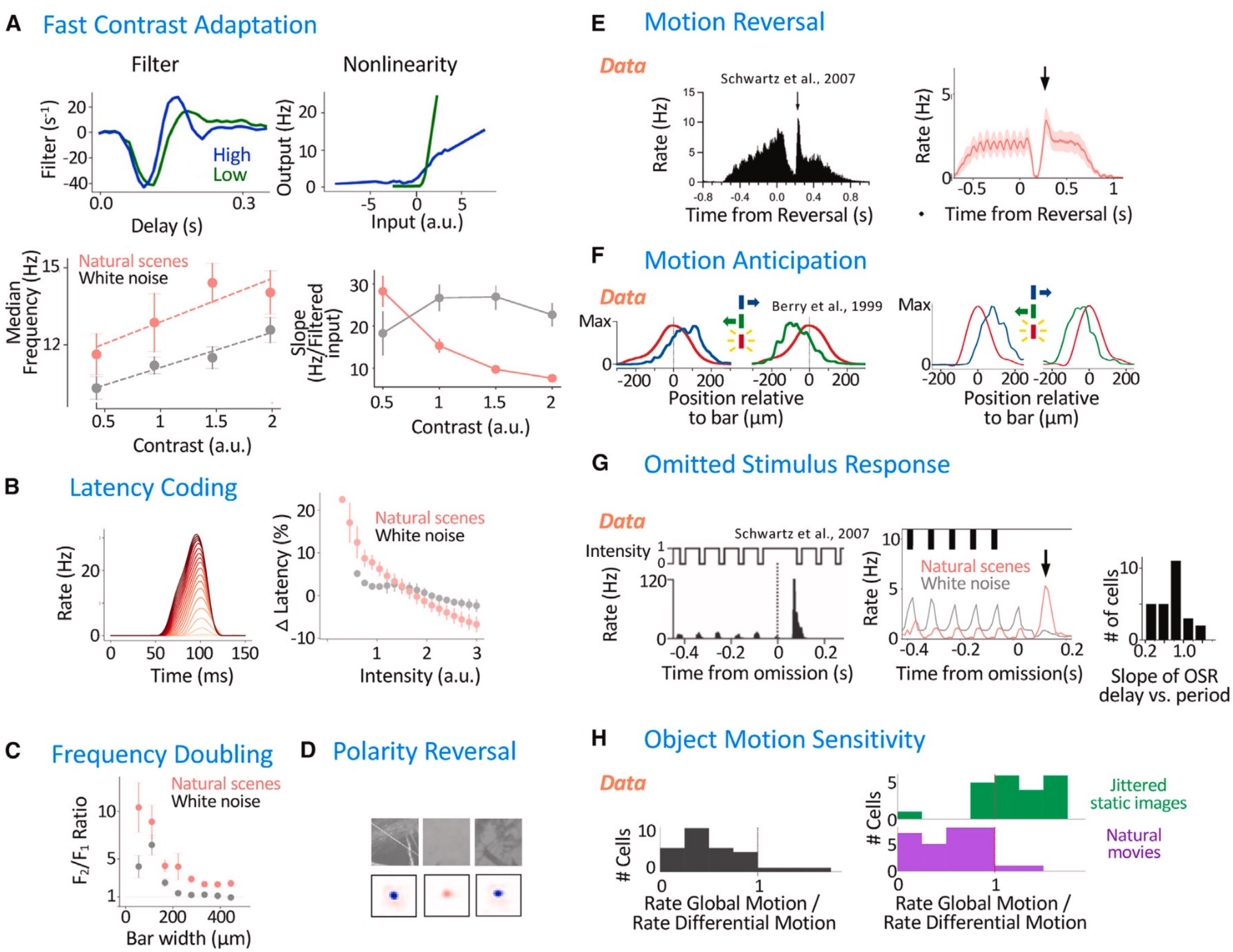




Our deep network model of the retina predicts the retinal response to natural movies almost as well as can be expected given intrinsic stochasticity in the retina itself.

Hidden units of our model retina behave like hidden units of the biological retina.

We performed 8 seminal experiments spanning > 2 decades on our model retina. It behaved exactly like the biological retina on all 8 experiments.

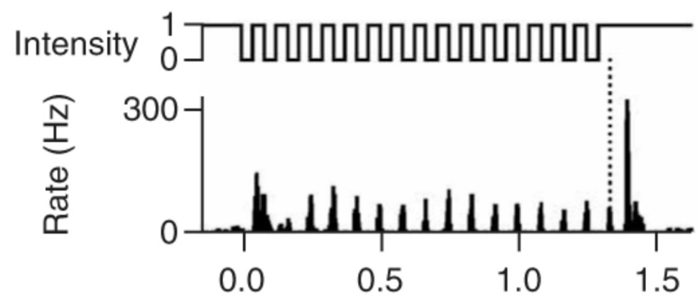


CNNs trained on natural scenes, but not white noise, generalize to artificial stimuli

The retina as a predictive engine

The omitted stimulus response:
Predicting periodic signals

Schwartz et. al. 2007

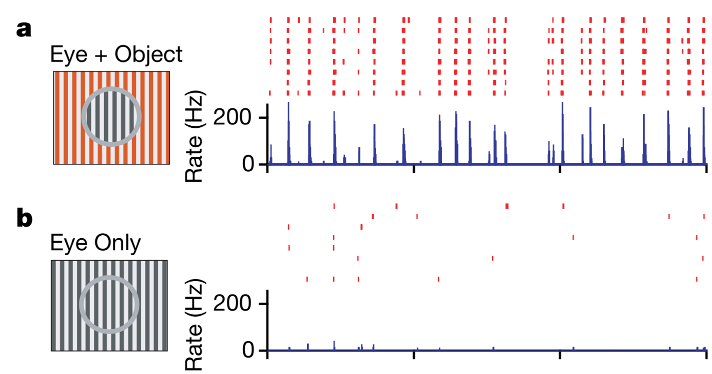


Motion reversal response:
Predicting future motion

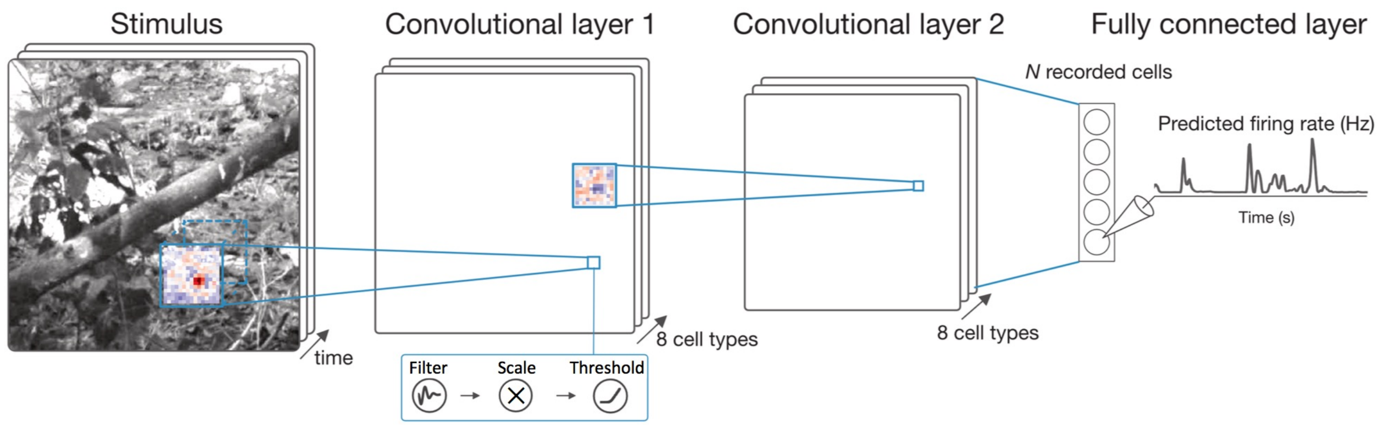
Schwartz et. al. 2007

Segregating moving objects
from background motion:
Filtering out predictions from eye movements

Olveczky, Baccus, Meister, 2007

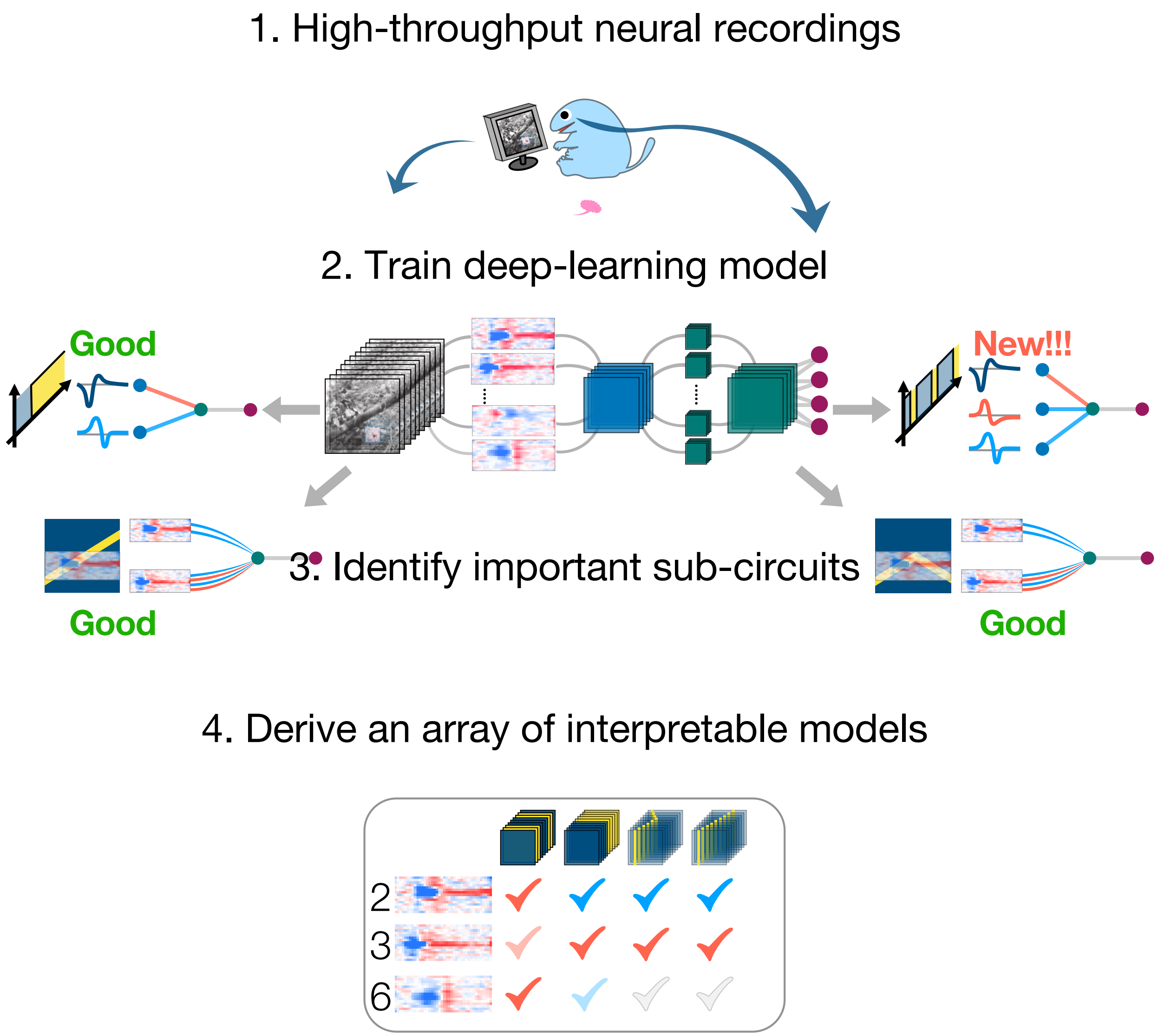


A case study: machine learning and the retina



Raises deep questions about the nature of explanation in neuroscience.

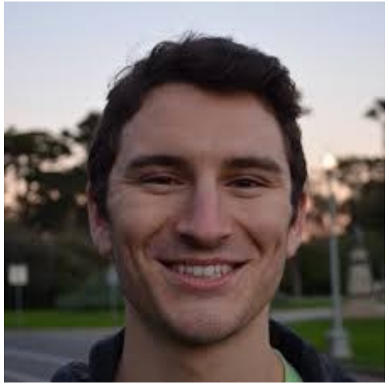
- But we can understand at least this model – and how it simultaneously accounts for all the experiments.
- Qualitative phenomena can yield qualitative explanations.
- Intriguingly – once we have become quantitative, we have to step back and become qualitative to obtain conceptual understanding.
- Next :A systematic approach to understanding through model reduction

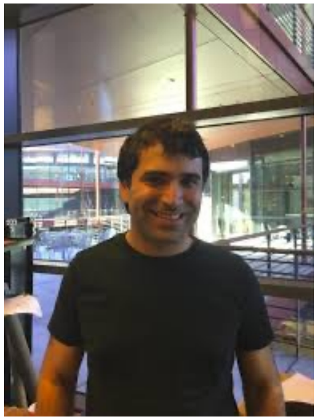

1. High-throughput neural recordings
2. Train deep-learning model
Good
New!!!
3. Identify important sub-circuits
Good
Good
4. Derive an array of interpretable models
2
3
6

Mechanistic interpretability for grid cells

B. Sorscher*, G. Mel*, S. Ganguli, S. Ocko, A unified theory for the origin of grid cells through the lens of pattern formation, NeurIPS 2019.

B. Sorscher*, G. Mel*, S. Ocko, L. Giocomo, S. Ganguli, A unified theory for the computational and mechanistic origins of grid cells Neuron 2023.





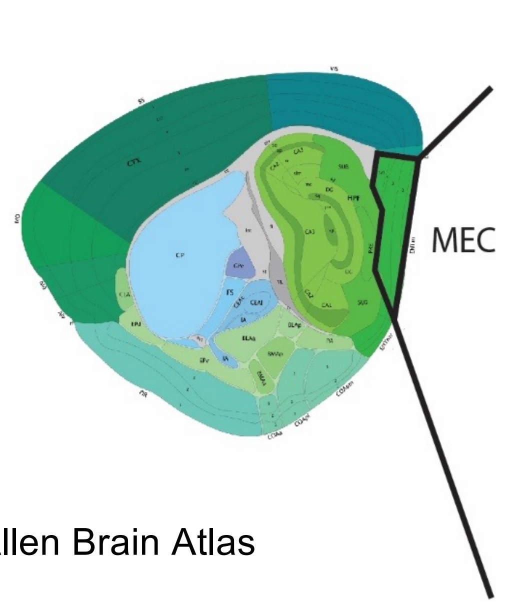
Ben Sorscher*

Gabriel Mel*

Sam Ocko

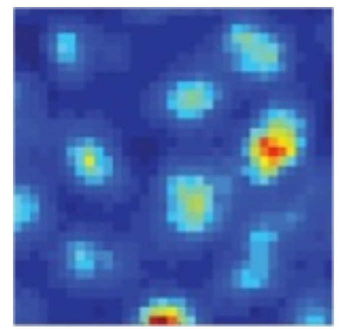
* equal contribution

What is the function of MEC?
What role might grid cells play in this function?



Allen Brain Atlas

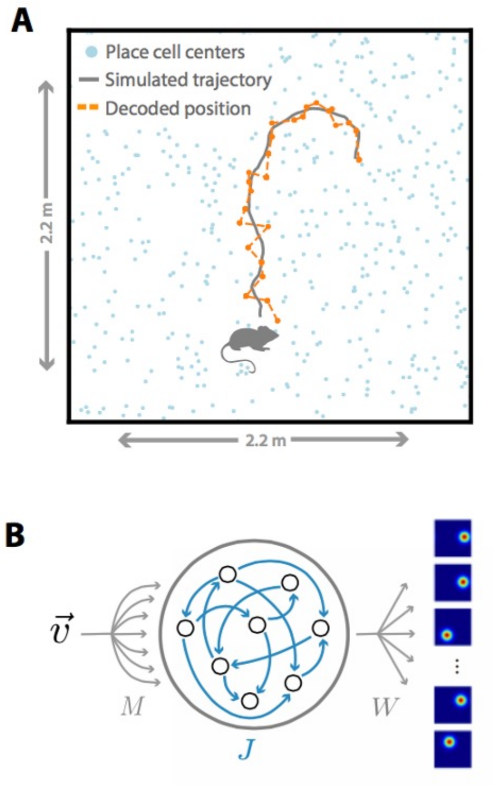
Grid Cell



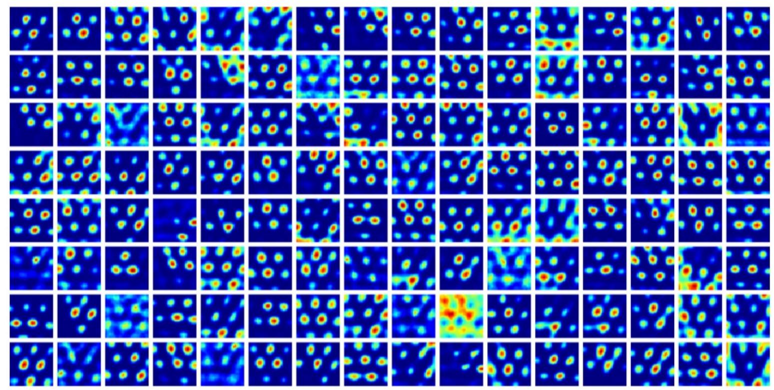
Why did evolution order grid cells?

Is it a consequence of solving some task under some simple biological constraints?

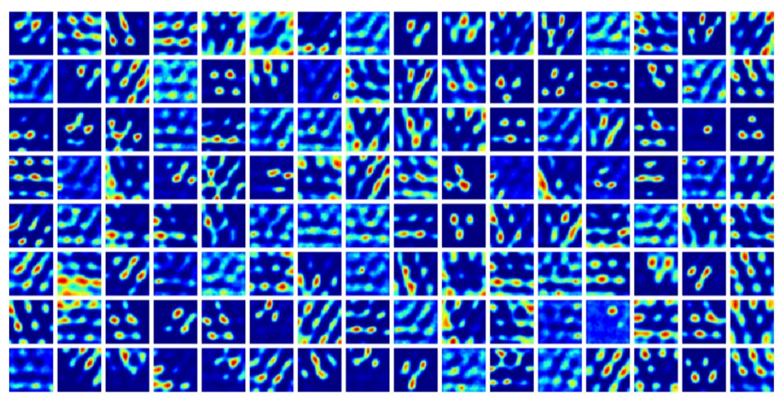
Nonnegative firing rates + center-surround place cells lead to hexagonal grids



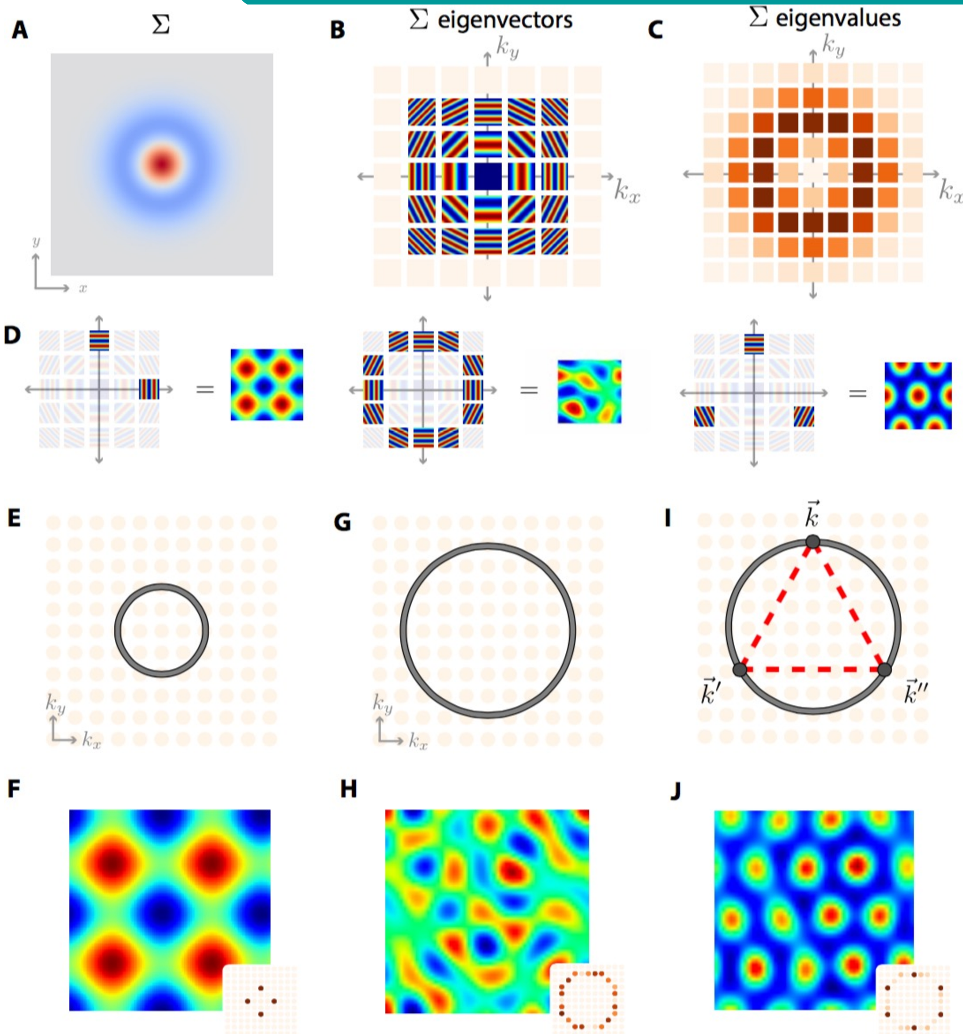
Grid scores 1.42 -- 1.18



Grid scores 0.41 -- 0.38



A mathematical theory for emergent lattices in neural nets



B. Sorscher*, G. Mel*, S. Ganguli, S. Ocko, A unified theory for the origin of grid cells through the lens of pattern formation, NeurIPS 2019.

B. Sorscher*, G. Mel*, S. Ocko, L. Giocomo, S. Ganguli, A unified theory for the computational and mechanistic origins of grid cells Neuron 2023.

Interpretability and control in AI

Locating and Editing Factual Associations in GPT

Kevin Meng* MIT CSAIL
David Bau* Northeastern University
Alex Andonian MIT CSAIL
Yonatan Belinkov† Technion – IIT

Approximate Data Deletion from Machine Learning Models

Zachary Izzo Dept. of Mathematics, Stanford University, zizzo@stanford.edu
Mary Anne Smart Department of CS&E, UC San Diego, msmart@eng.ucsd.edu
Kamalika Chaudhuri Department of CS&E, UC San Diego, kamalika@cs.ucsd.edu
James Zou Deptartment of BDS, Stanford University, jamesz@stanford.edu

A Mathematical Framework for Transformer Circuits

AUTHORS

Nelson Elhage*†, Neel Nanda*, Catherine Olsson*, Tom Henighan†, Nicholas Joseph†, Ben Mann†, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, Chris Olah‡

AFFILIATION

Anthropic

PUBLISHED

Dec 22, 2021

* Core Research Contributor; † Core Infrastructure Contributor; ‡ Correspondence to colah@anthropic.com; Author contributions statement below.

Why you saw this ad

Only you can see this

Different factors contribute to why you saw an ad, including who the advertiser chooses to reach and your activity. These factors help us predict ads you might like.

Advertiser choices
View who Solar-Estimate.org wants to show ads to

Your activity
You interacted with content about home services and more

Learn more about other factors that contribute to our ad predictions.

What you can do

Hide all ads from this advertiser — Hide

Update your ad preferences

Was this information useful? Yes No

Outline: today's frontiers and tomorrow challenges/opportunities

(1) Alternate physical implementations of ML algorithms: negotiating high dimensional error landscapes with atoms and photons in open dissipative open systems.

(2) The physical origins of generative AI: from diffusion models to societal implications.

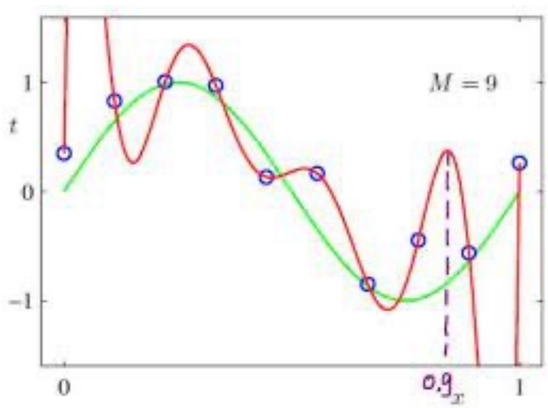
(3) Mechanistic interpretability: a unification of neuroscience, physics and AI with applications to alignment and government regulation

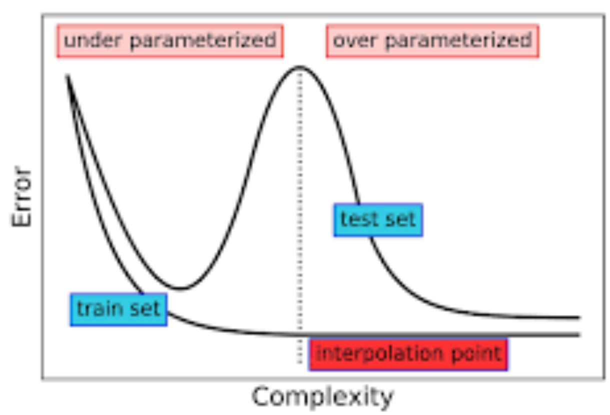
(4) Theory: the generalization puzzle: how can huge networks generalize without overfitting.

(5) Unsustainable neural scaling laws and the combatting the data hungriness of modern AI

(6) The mystery of structured sequences: the unreasonable effectiveness of predicting every next word on the internet (+ human feedback). LLMs as the next challenge/opportunity for theory and society.

Classical dogma to prevent overfitting





Data generated from a simple function (low degree polynomial)

Fit data using a complex function (high degree polynomial)

With M data points an order M polynomial will go exactly through all the data points -> training error zero.

But it will oscillate wildly between data points -> test error high.

But with M data points lets fit a degree P>>M polynomial via gradient descent starting with all coefficients 0.

There are many many degree P polynomials with P>>M that achieve training error 0 and high test error.

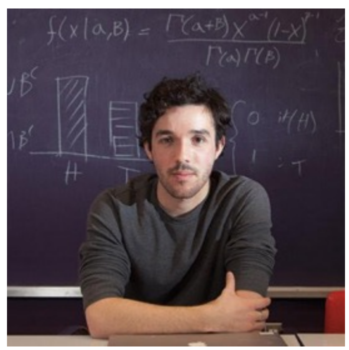
But gradient descent starting from 0 coefficients will pick a “simple” polynomial amongst all these that does not overfit.

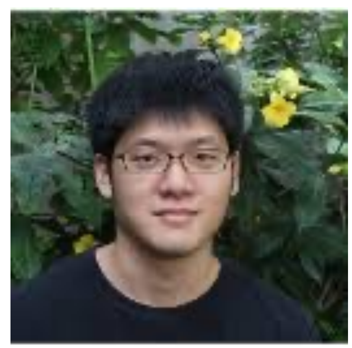
Gradient descent has an implicit bias when the space of zero training error solutions is large.

An implicit bias of SGD: stochastic collapse to invariant sets corresponding to simpler subnetworks

Stochastic Collapse: How Gradient Noise Attracts SGD Dynamics Towards Simpler Subnetworks

Feng Chen* **Daniel Kunin*** **Atsushi Yamamura (山村篤志)*** **Surya Ganguli**
Stanford University
{fengc,kunin,atsushi3,sganguli}@stanford.edu





An implicit bias of SGD: stochastic collapse to invariant sets corresponding to simpler subnetworks

Question: Why do overparameterized deep neural networks generalize?

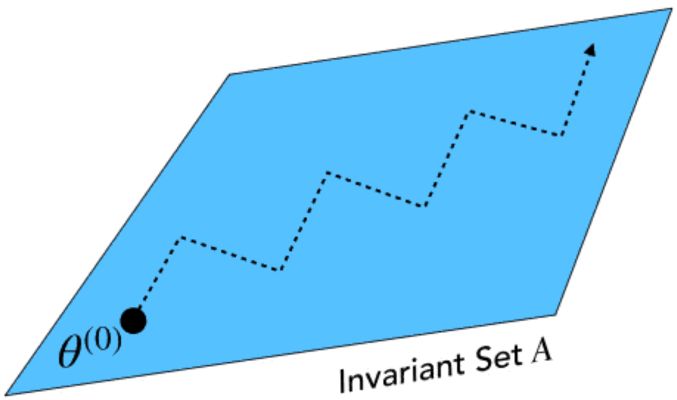
Perhaps SGD has implicit biases that restrict it to exploring a limited regime of parameter space?

Many works on this topic.

We provide a new, additional implicit bias driven by symmetry that applies very generally to many different deep networks that are not toy models.

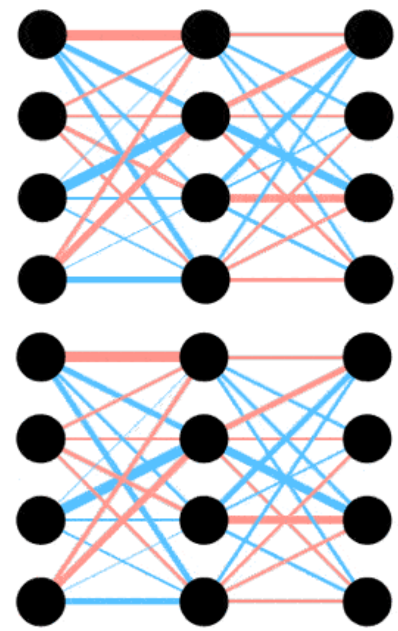
Many invariant sets exist that can trap SGD dynamics

Invariant Set: *A subspace of parameters* $A \subseteq \mathbb{R}^d$ *such that given any initialization* $\theta^{(0)} \in A$ *all future iterates of SGD* $\theta^{(t)}$ *for* $t \geq 0$ *are contained within* A*, for any batch size, learning rate, and mini-batches.*



Sign Invariant Set
(vanishing neurons)

Permutation Invariant Set
(identical neurons)



Once SGD enters an invariant set, it stays there forever.
But does it ever get close to any of them?
Yes, because of the statistical structure of SGD noise.

Stochastic gradient descent and stochastic gradient flow as a model: the importance of position dependent diffusion

SGD: $$\theta^{(t+1)} = \theta^{(t)} - \frac{\eta}{\beta} \sum_{i \in \mathcal{B}^{(t)}} \nabla_\theta \ell \left(\theta^{(t)}; x_i, y_i \right)$$

$\theta =$ parameters

$\eta =$ learning rate

$\beta =$ batch size

Every gradient w.r.t every example lies within the invariant subspace. Proj to orthog complement is 0.

Stochastic gradient descent and stochastic gradient flow as a model: the importance of position dependent diffusion

θ = parameters

η = learning rate

β = batch size

SGD: $$\theta^{(t+1)} = \theta^{(t)} - \frac{\eta}{\beta} \sum_{i \in \mathcal{B}^{(t)}} \nabla_\theta \ell \left(\theta^{(t)}; x_i, y_i \right)$$

Every gradient w.r.t every example lies within the invariant subspace. Proj to orthog complement is 0.

SGF: $$d\theta_t = -\nabla \mathcal{L}(\theta_t) dt + \sqrt{\frac{\eta}{\beta}} \Sigma(\theta_t) dB_t, \qquad \theta_0 = \theta^{(0)}$$

$$\Sigma(\theta_t) = \sqrt{\mathbb{E}[\xi_{\mathcal{B}}(\theta)\xi_{\mathcal{B}}(\theta)^\intercal]}$$

$$D(\theta_t) = \frac{\eta}{2\beta}\Sigma(\theta_t)\Sigma(\theta_t)^\intercal$$

Position dependent diffusion constant!

$$P^\perp D(\theta_t) P^\perp = 0 \text{ for } \theta_t \text{ in invariant subspace}$$

Diffusion in directions perpendicular to an invariant subspace vanishes!

Stochastic gradient descent and stochastic gradient flow as a model: the importance of position dependent diffusion

$\theta = \text{parameters}$

$\eta = \text{learning rate}$

$\beta = \text{batch size}$

SGD: $$\theta^{(t+1)} = \theta^{(t)} - \frac{\eta}{\beta} \sum_{i \in \mathcal{B}^{(t)}} \nabla_\theta \ell \left(\theta^{(t)}; x_i, y_i \right)$$

Every gradient w.r.t every example lies within the invariant subspace. Proj to orthog complement is 0.

SGF: $$d\theta_t = -\nabla \mathcal{L}(\theta_t) dt + \sqrt{\frac{\eta}{\beta}} \Sigma(\theta_t) dB_t, \qquad \theta_0 = \theta^{(0)}$$

$$\Sigma(\theta_t) = \sqrt{\mathbb{E}[\xi_{\mathcal{B}}(\theta)\xi_{\mathcal{B}}(\theta)^\intercal]}$$

$$D(\theta_t) = \tfrac{\eta}{2\beta}\Sigma(\theta_t)\Sigma(\theta_t)^\intercal$$

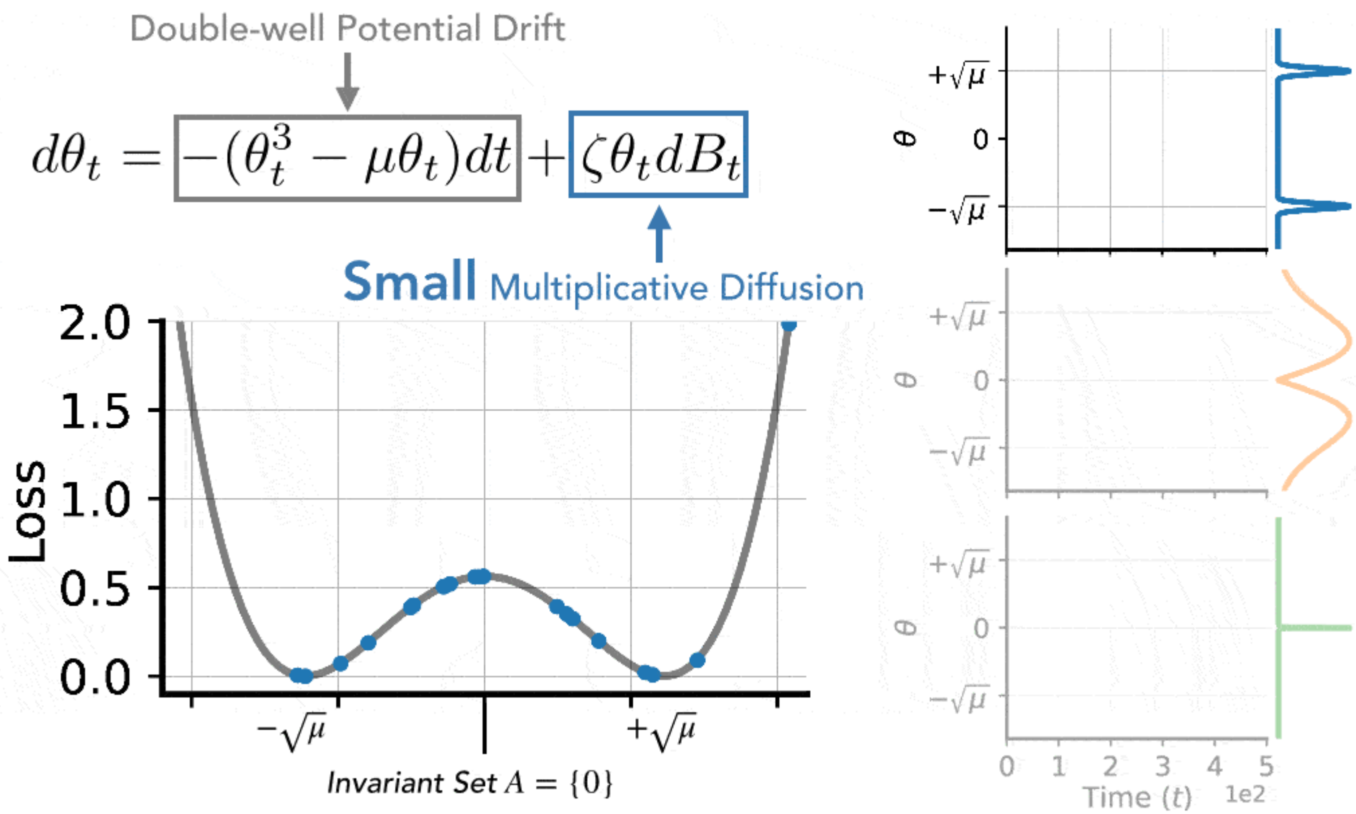
Position dependent diffusion constant!

$$P^\perp D(\theta_t) P^\perp = 0 \text{ for } \theta_t \text{ in invariant subspace}$$

Diffusion in directions perpendicular to an invariant subspace vanishes!

Lemma: invariant subspaces of SGD are also invariant subspaces of SGF.

An invariant set can be attractive: “stochastic collapse”



Larger noise makes invariant set more attractive

When will SGF be attracted to an invariant set?
A local analysis

A definition from stochastic control theory:

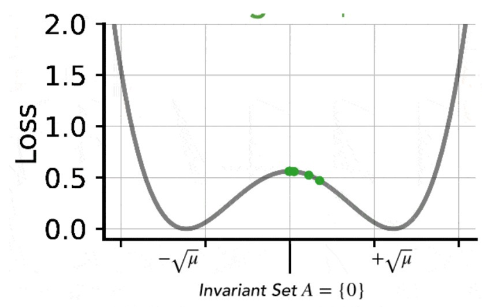
Definition 4.1 (Stochastic Attractivity). *An invariant set $A \subset \mathbb{R}^d$ of a stochastic process $\{\theta_t \in \mathbb{R}^d : t \geq 0\}$ is stochastically attractive*[6] *if for any $\rho > 0$ and $\epsilon > 0$, there exists $\delta > 0$ such that for any $\theta \in \mathbb{R}^d$ with $d(\theta, A) < \delta$,*

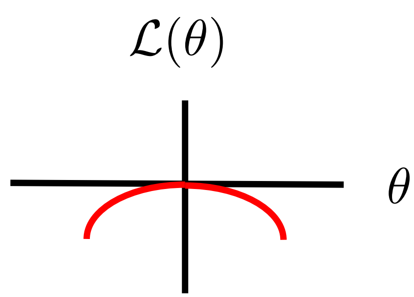
$$\mathbb{P}\left[\sup_{t\geq 0} d(\theta_t, A) \geq \epsilon \middle| \theta_0 = \theta\right] \leq \rho, \tag{4}$$

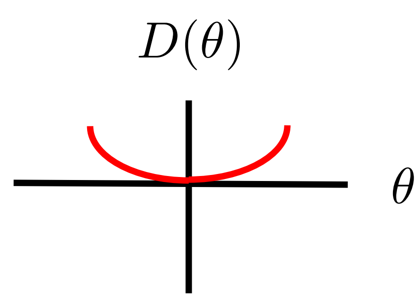
where $d(\theta, A)$ is the Euclidian distance between θ and A.

Informally: if you want all future iterates to remain close to the set with high probability, there always Exists an initial condition that achieves that.

When will SGF be attracted to an invariant set?
A local analysis







$\partial_\theta^2 \mathcal{L}|_{\theta=0}$ $\partial_\theta^2 \mathcal{D}|_{\theta=0}$

Origin is stochastically attractive iff $\partial_\theta^2 \mathcal{L}|_{\theta=0} + \partial_\theta^2 \mathcal{D}|_{\theta=0} > 0$

Note: can stochastically collapse to maxima or saddle points more generally!

When will SGF be attracted to an invariant set? A local analysis: sufficient condition in multiple dimensions

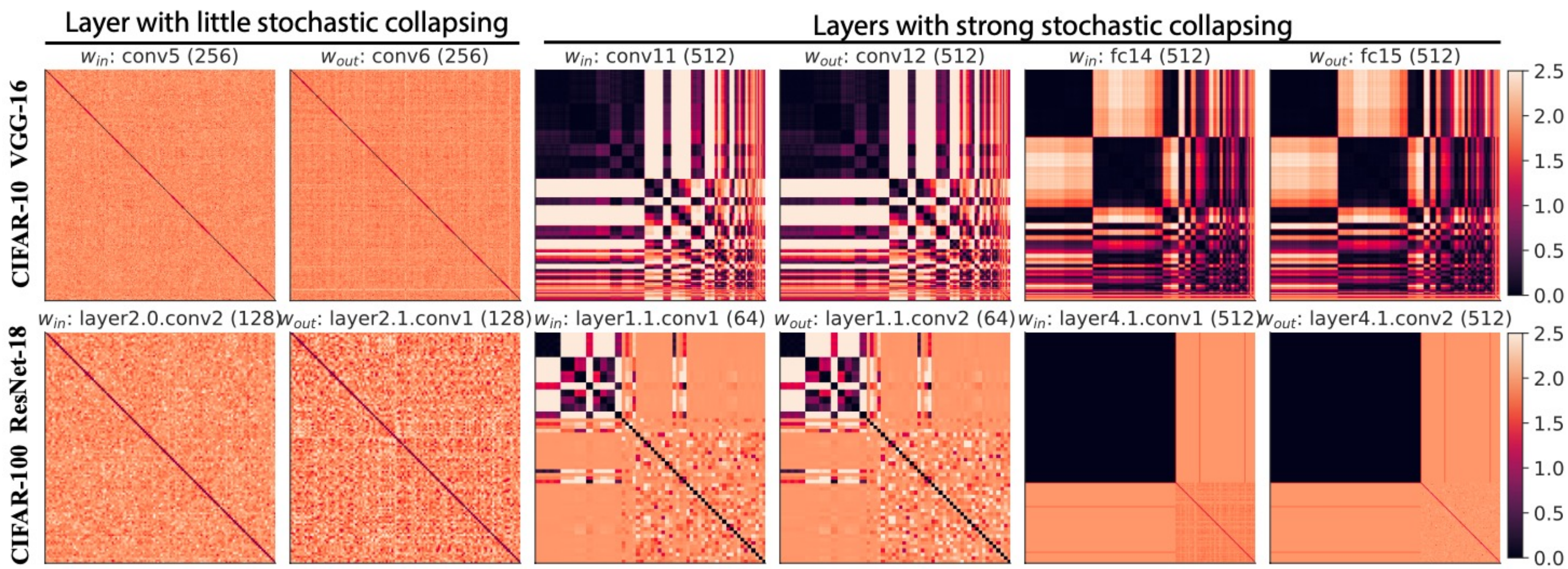
Theorem 4.2 (A Sufficient Condition for Stochastic Attraction in High-Dimensions). *Let $A \subset \mathbb{R}^d$ be a d_A-dimensional affine subset, and a stochastic process $\{\theta_t \in \mathbb{R}^d : t \geq 0\}$ obey Eq. 3 in A_c, open c-neighborhood of A with some $c > 0$. Suppose $\mathcal{L} : A_c \to \mathbb{R}$ is a C^3-function whose first and second-order derivatives are L-Lipschitz continuous. $D : A_c \to \mathbb{R}^{d \times d}$ is the diffusion matrix such that the second-order derivatives of its elements are L-Lipschitz continuous. Furthermore, we assume that all the elements of $\sqrt{D} : A_c \to \mathbb{R}^{d \times d}$ are L-Lipschitz continous. Let $D_\perp = P_\perp D P_\perp$ where $P_\perp : \mathbb{R}^d \to \mathbb{R}^d$ projects to the normal space of A. If there exists $\delta > 0$ such that $\nabla^2_{\hat{n}} \hat{n}^\intercal D \hat{n}^\intercal > \delta$ and*

$$\nabla^2_{\hat{n}} \left(\mathcal{L} - \frac{1}{2} \operatorname{Tr} D_\perp \; + (1 - \delta) \hat{n}^\intercal D \hat{n} \right) > 0, \tag{6}$$

for any unit normal vector $\hat{n} \in \mathbb{R}^d$ perpendicular to A and $\theta \in A$, then A is stochastically attractive.

Intuition: diffusion tensor in directions perpendicular to invariant subspace must grow sufficiently quickly to counter act drift terms perpendicular to subspace and away from it

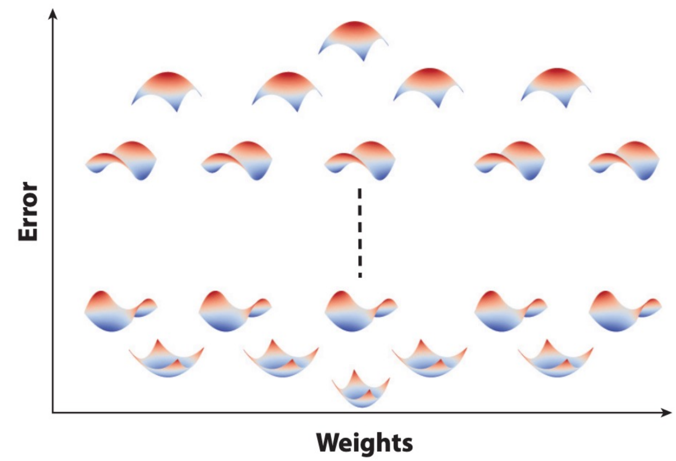
Stochastic collapse in the wild can be prominent



Hierarchically cluster incoming weights into all neurons (channels) in a given layer
Plot the outgoing weights using the same clustering order derived from incoming weights.

Later layers are larger than earlier layers; these later overparameterized layers show collapse.

Error landscape perspective: higher training error saddle points can achieve lower test error than lower training error minima



Error landscape perspective:
There are many (manifolds) of critical points (saddle points and minima)

The best generalizer is a saddle! It may have higher training error than local minima, but it will have lower test error.

Local minima will have lower training error, but higher test error.

Stochastic collapse gets attract to good generalizing saddles and avoids local minima through position dependent diffusion dynamics.

Outline: today's frontiers and tomorrow challenges/opportunities

(1) Alternate physical implementations of ML algorithms: negotiating high dimensional error landscapes with atoms and photons in open dissipative open systems.

(2) The physical origins of generative AI: from diffusion models to societal implications.

(3) Mechanistic interpretability: a unification of neuroscience, physics and AI with applications to alignment and government regulation

(4) Theory: the generalization puzzle: how can huge networks generalize without overfitting.

(5) Unsustainable neural scaling laws and the combatting the data hungriness of modern AI

(6) The mystery of structured sequences: the unreasonable effectiveness of predicting every next word on the internet (+ human feedback). LLMs as the next challenge/opportunity for theory and society.

Humans are still much more data efficient than AI

If humans read as much as GPT3 did it would take 20,000 years.

If humans practiced GO as much as Alpha-Go did we would have to play 300 games a day for 30 years.

Training with less data is essential: neural scaling laws are slow and unsustainable

Neural scaling laws: error tends to fall off as a power law with: the number of parameters,
the number of data points,
or amount of compute

Training with less data is essential: neural scaling laws are slow and unsustainable

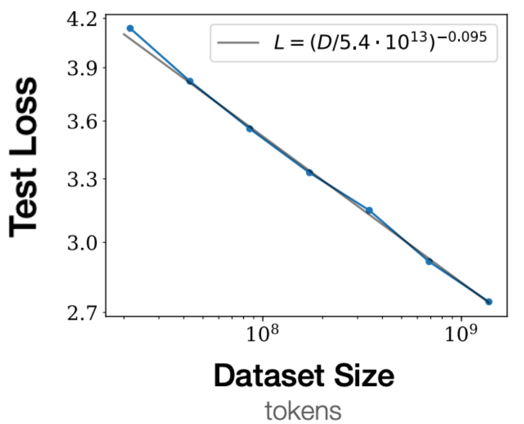
Neural scaling laws: error tends to fall off as a power law with: the number of parameters,
the number of data points,
or amount of compute

But: the exponent is small, so advancing AI through scaling alone is expensive and unsustainable.

Training with less data is essential: neural scaling laws are slow and unsustainable

Neural scaling laws: error tends to fall off as a power law with: the number of parameters, the number of data points, or amount of compute

But: the exponent is small, so advancing AI through scaling alone is expensive and unsustainable.



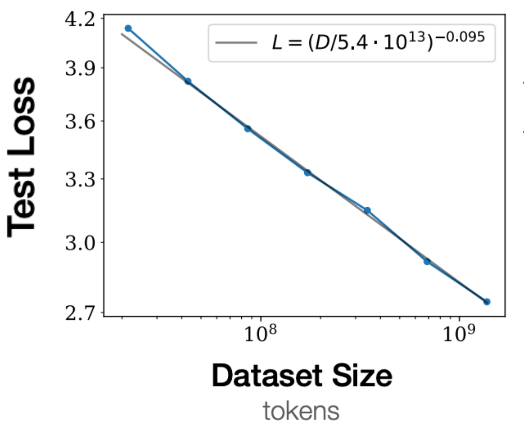
Language modelling w/ large transformers:
Drop in cross-entropy from 3.4 nats to 2.8 nats
Requires 10x more training data

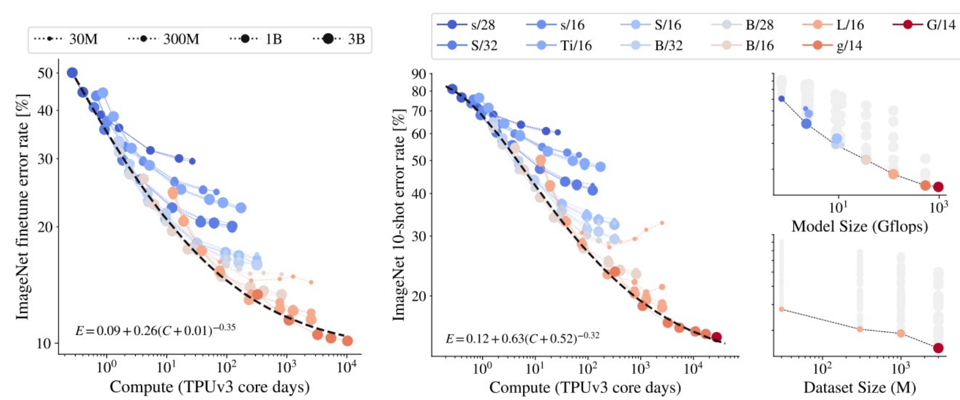
Kaplan et. al. arxiv: 2001.08361

Training with less data is essential: neural scaling laws are slow and unsustainable

Neural scaling laws: error tends to fall off as a power law with: the number of parameters, the number of data points, or amount of compute

But: the exponent is small, so advancing AI through scaling alone is expensive and unsustainable.





Language modelling w/ large transformers:
Drop in cross-entropy from 3.4 nats to 2.8 nats
Requires 10x more training data

Kaplan et. al. arxiv: 2001.08361

Large vision transformers: going from 1B to 2B datapoints
Leads to a drop of ~2 to 3 percent error on ImageNet.

Zhai et. al. arxiv: 2106.04560

Beating neural scaling laws

Ben Sorscher

Robert Geirhos

Amro Abbas

Mansheej Paul

Karolina Dziugaite

Ari Morcos

Beyond neural scaling laws:
beating power law scaling via data pruning

Ben Sorscher[*1] **Robert Geirhos**[*2] **Shashank Shekhar**[3]

Surya Ganguli[1,3§] **Ari S. Morcos**[3§]

[*]equal contribution, work done during an internship at Meta AI
[1]Department of Applied Physics, Stanford University
[2]University of Tübingen
[3]Meta AI
[§]Joint senior authors

NeurIPS 2022
Outstanding paper award

Deep Learning on a Data Diet:
Finding Important Examples Early in Training

Mansheej Paul
Stanford University
mansheej@stanford.edu

Surya Ganguli
Stanford University; Facebook AI Research
sganguli@stanford.edu

Gintare Karolina Dziugaite
Mila *
gkdz@google.com

NeurIPS 2021

SemDeDup: Data-efficient learning at web-scale
through semantic deduplication

Arxiv: 2303.09543

Amro Abbas[1] **Kushal Tirumala**[1*] **Dániel Simig**[1*] **Surya Ganguli**[2] **Ari S. Morcos**[1*]
[1]Meta AI (FAIR) [2]Department of Applied Physics, Stanford University

Key ideas and take home messages

- Slow power law scaling of error with data set size indicates training examples are highly redundant.
- Therefore we should be able to prune datasets to identify sparse subsets of non-redundant examples.
- If we plot error as a function of the size of a pruned non-redundant dataset, we might be able to beat power law scaling.

Key ideas and take home messages

- Slow power law scaling of error with data set size indicates training examples are highly redundant.
- Therefore we should be able to prune datasets to identify sparse subsets of non-redundant examples.
- If we plot error as a function of the size of a pruned non-redundant dataset, we might be able to beat power law scaling.
- We show analytically we can do this in theory for perceptron learning in a student-teacher setting (achieving at least exponential scaling!)
- We show in practice how to beat power law scaling for:
  ResNets trained on SVHN, CIFAR10, ImageNet
  Vision Transformers pre-trained on ImageNet & fine-tuned on CIFAR10
- We perform a benchmarking study of 8 different pruning metrics + a new metric we develop that is

Asymptotically exact calculation of test error for perceptron learning with non-Gaussian pruned data

- Important parameters of the problem:

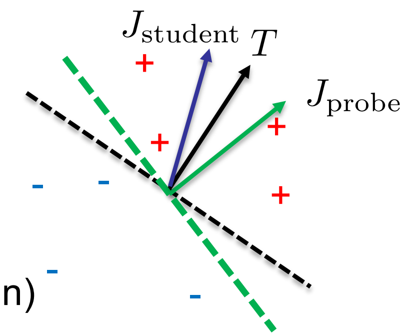
  θ = angle between T and J_{probe}

  z = projection of data x onto J_{probe}

  p(z) = distribution of data along z (non-Gaussian)

  κ = margin of $J_{student}$ w.r.t. training data

  α = P/N = ratio of examples to parameters



- Quantities which concentrate in the high dimensional statistics limit

  ρ = cosine angle between $J_{student}$ and teacher J

  R = cosine angle between $J_{student}$ and teacher T

  Test error $\varepsilon = \cos^{-1}(R)/\pi$

Self-consistent replica equations relating R and ρ to θ, p(z), κ and α.

$$\frac{R-\rho\cos\theta}{\sin^2\theta} = \frac{\alpha}{\pi\Lambda}\left\langle \int_{-\infty}^{\kappa} dt\ \exp\left(-\frac{\Delta(t,z)}{2\Lambda^2}\right)(\kappa-t)\right\rangle_z \quad (1)$$

$$1-\frac{\rho^2+R^2-2\rho R\cos\theta}{\sin^2\theta} = 2\alpha\left\langle \int_{-\infty}^{\kappa} dt\, \frac{e^{-\frac{(t-\rho z)^2}{2(1-\rho^2)}}}{\sqrt{2\pi}\sqrt{1-\rho^2}} H\left(\frac{\Gamma(t,z)}{\sqrt{1-\rho^2}\Lambda}\right)(\kappa-t)^2\right\rangle_z \quad (2)$$

$$\frac{\rho-R\cos\theta}{\sin^2\theta} = 2\alpha\left\langle \int_{-\infty}^{\kappa} dt\, \frac{e^{-\frac{(t-\rho z)^2}{2(1-\rho^2)}}}{\sqrt{2\pi}\sqrt{1-\rho^2}} H\left(\frac{\Gamma(t,z)}{\sqrt{1-\rho^2}\Lambda}\right)\left(\frac{z-\rho t}{1-\rho^2}\right)(\kappa-t) + \frac{1}{2\pi\Lambda}\exp\left(-\frac{\Delta(t,z)}{2\Lambda^2}\right)\left(\frac{\rho R-\cos\theta}{1-\rho^2}\right)(\kappa-t)\right\rangle_z \quad (3)$$

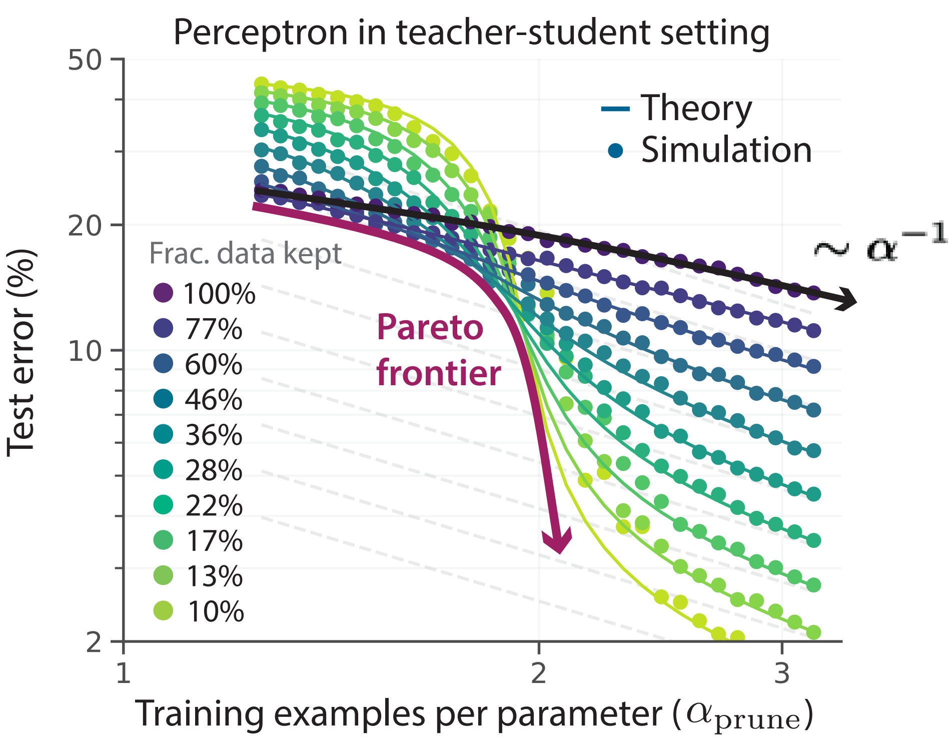
Where,

$$\Lambda = \sqrt{\sin^2\theta - R^2 - \rho^2 + 2\rho R\cos\theta}, \quad (4)$$

$$\Gamma(t,z) = z(\rho R - \cos\theta) - t(R-\rho\cos\theta), \quad (5)$$

$$\Delta(t,z) = z^2\left(\rho^2+\cos^2\theta - 2\rho R\cos\theta\right) + 2tz(R\cos\theta-\rho) + t^2\sin^2\theta. \quad (6)$$

Towards a mathematical theory of data pruning: the student-teacher perceptron setting

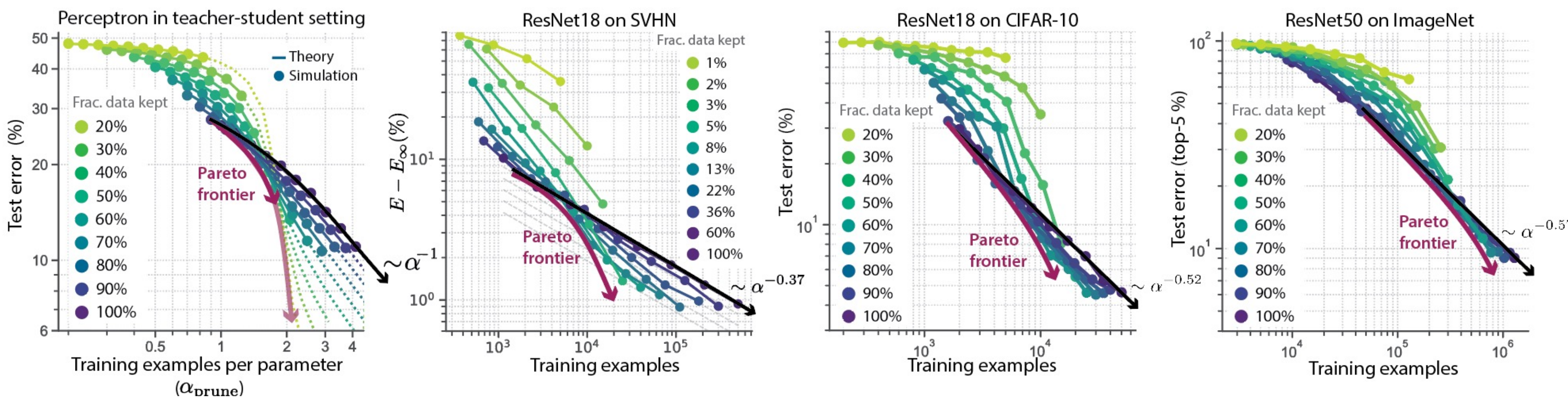


$$\alpha_{\text{prune}} = f\alpha_{\text{total}}$$

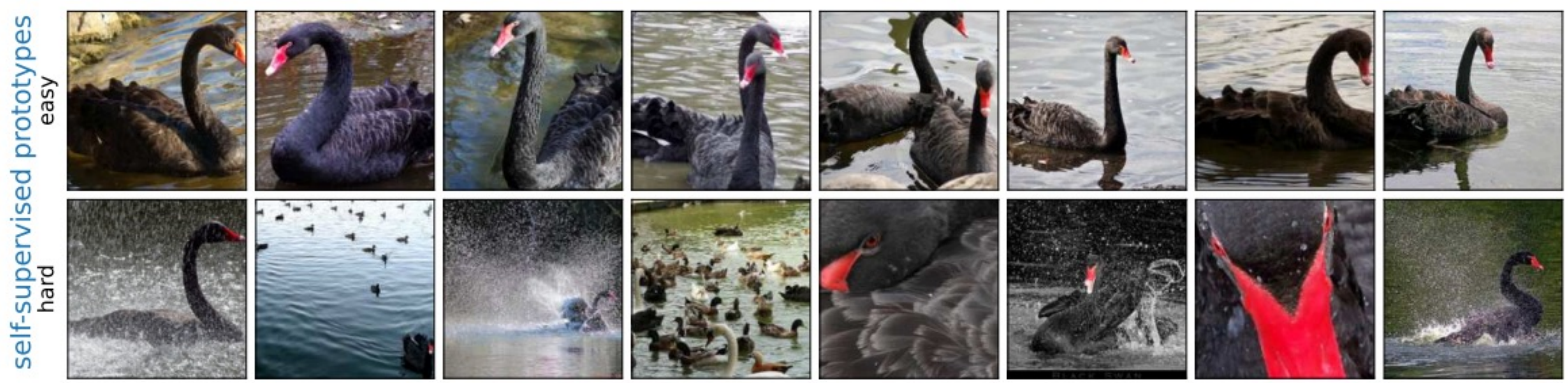
Pruning with a perfect metric, according to the margin on the teacher's decision boundary (θ=0)

- At any fixed pruning strategy with a fixed f, as one increases α_{total} and therefore α_{prune}, test error falls off as a power law.
- However at any fixed α_{prune}, if one finds a Pareto-optimal tradeoff between α_{total} and f then Pareto frontier yields at least exponential decay of test error with α_{prune}.
- We find that at larger initial dataset size α_{total} , more aggressive pruning (smaller f) yields better performance at fixed α_{prune}.

Beating power law scaling in practice



Examples of hard and easy images



Imagenet class 100 (black swan)

Examples of hard and easy images

Imagenet class 200 (Tibetan terrier)

Examples of hard and easy images

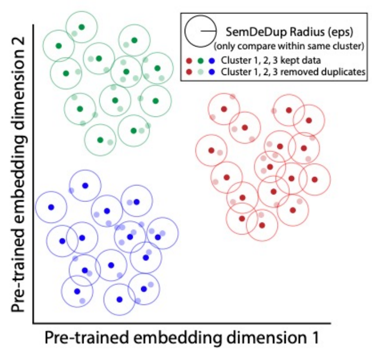
ImageNet class 300 (Tiger beetle)

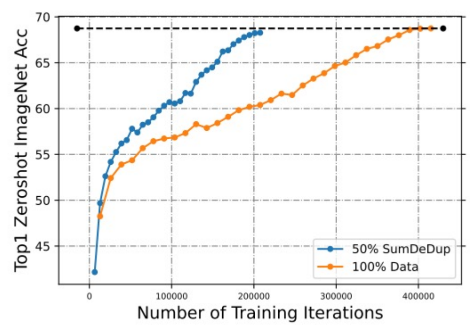
Scaling up to webscale

SemDeDup: Data-efficient learning at web-scale through semantic deduplication

Arxiv: 2303.09543

Amro Abbas[1] **Kushal Tirumala**[1*] **Dániel Simig**[1*] **Surya Ganguli**[2] **Ari S. Morcos**[1*]
[1]Meta AI (FAIR) [2]Department of Applied Physics, Stanford University





Training a CLIP model on a highly curated 440M subset of LAION.

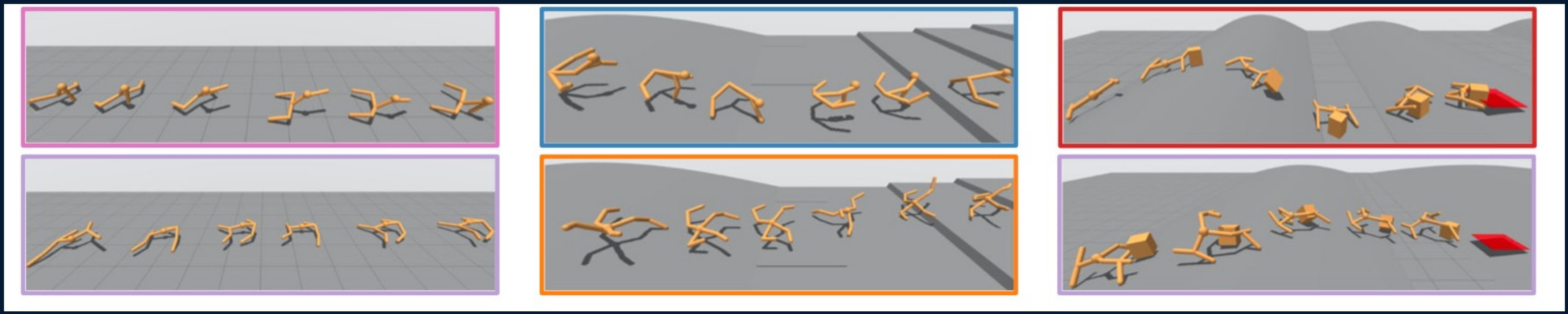
Can drop 50% of the data points down to 220M

And suffer no loss on ImageNet + ~20 other downstream tasks.

Data efficiency from learning across multiple time scales

Current AI:	Pretrain a foundation model on Large amounts of data.		Fine tune on a task using small amounts of data. Or in-context learning.	
Humans:	Biological Evolution	Cultural Evolution	Lifetime Learning	Task Learning

Speeding up learning over a lifetime via biological evolution over generations

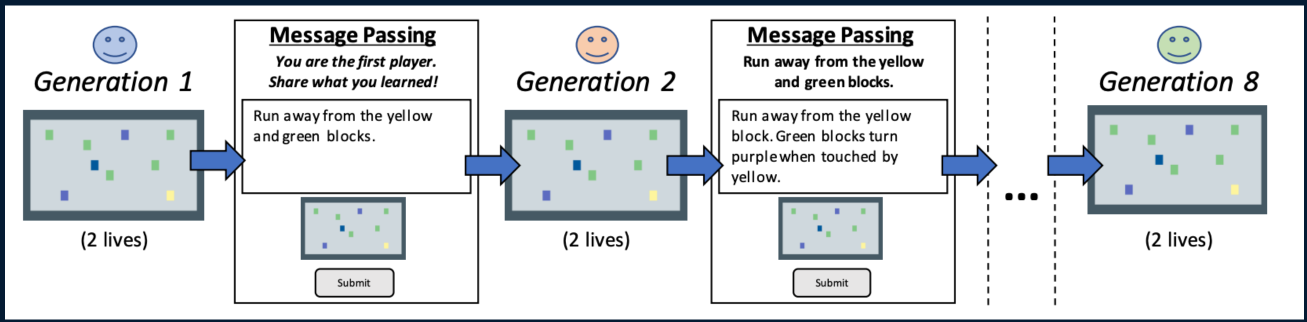


Only 10 generations of morphological evolution

lead to creatures that learned complex tasks more than twice as fast.

Gupta et. al. Nature Communications 2021 (Fei-Fei and Ganguli)

Characterizing human cultural learning

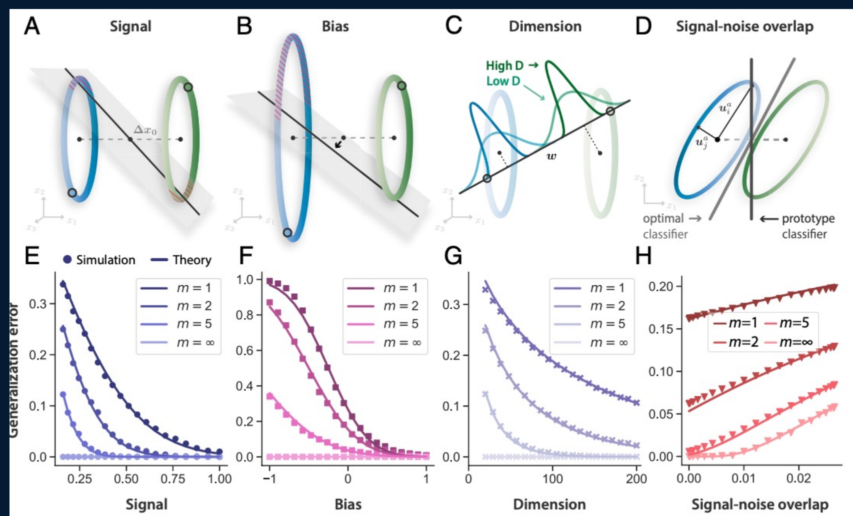


Humans can help later "generations" of humans learn faster

by communicating through language

Tessler et. al. NeurIPS 2021 (Goodman)

Data efficiency from examining neural representations: one-shot learning

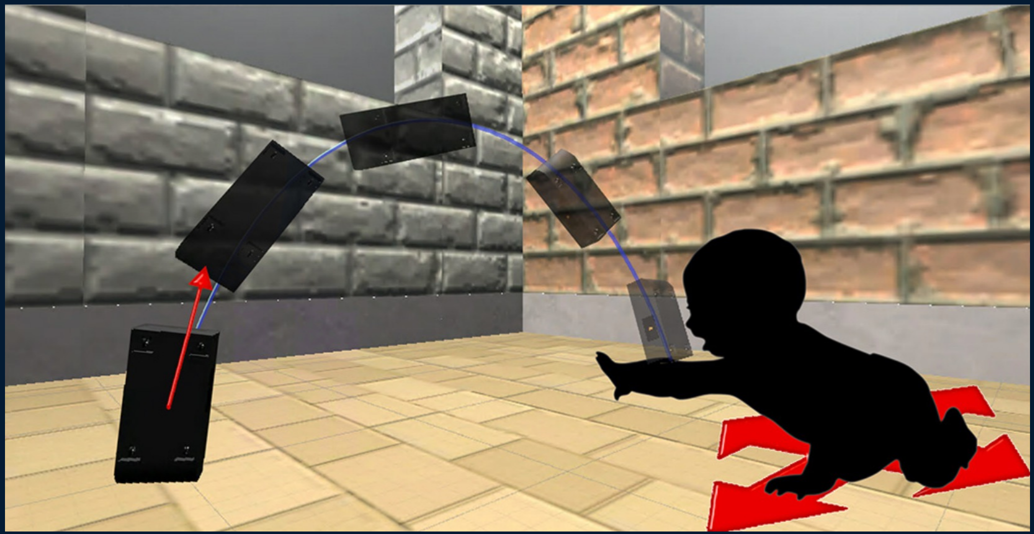


We developed a mathematical theory for how the geometry of neural representations can mediate one-shot learning of novel concepts.

We tested and verified our theory in both monkeys and machines, finding interesting differences.

Sorscher, Ganguli, Sompolinsky, PNAS 2022

Data efficiency from incorporating human like exploration



Mathematically formalized the notion of curiousity and instantiated it in artificial agents.

This lead to the creation of more structure behaviors which in turn

Could support more efficient reinforcement learning via exploration.

Haber et. al. Neurips 2018 (Fei-Fei and Yamins)

Outline: today's frontiers and tomorrow challenges/opportunities

(1) Alternate physical implementations of ML algorithms: negotiating high dimensional error landscapes with atoms and photons in open dissipative open systems.

(2) The physical origins of generative AI: from diffusion models to societal implications.

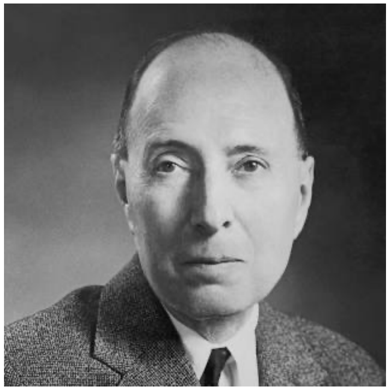
(3) Mechanistic interpretability: a unification of neuroscience, physics and AI with applications to alignment and government regulation

(4) Theory: the generalization puzzle: how can huge networks generalize without overfitting.

(5) Unsustainable neural scaling laws and the combatting the data hungriness of modern AI

(6) The mystery of structured sequences: the unreasonable effectiveness of predicting every next word on the internet (+ human feedback). LLMs as the next challenge/opportunity for theory and society.

The mystery of LLMs and structured sequences



The unreasonable effectiveness of mathematics in the natural sciences.

The unreasonable effectiveness of predicting every word on the internet.

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

The mystery of LLMs and structured sequences

SG

Provide a proof that there are an infinite number of primes.

One of the most famous proofs that there are an infinite number of primes was given by the ancient Greek mathematician Euclid, around 300 BCE. Here's a version of Euclid's proof:

Assume that there are only a finite number of prime numbers: p1, p2, p3, ..., pn.

Now, let's consider a new number Q, which is the product of all these primes plus one: Q = (p1 * p2 * p3 * ... * pn) + 1.

Now, Q is either a prime number or it isn't:

1. If Q is a prime number, then we've found a new prime that wasn't in our original list, which contradicts our assumption that there were only a finite number of primes.
2. If Q is not a prime number, then it must have at least one prime factor. But, Q cannot be divisible by any of the primes in our original list because when you divide Q by any of them, you get a remainder of 1 (due to the way we defined Q). So, Q must have a different prime factor that is not in our original list, which again contradicts our assumption that there were only a finite number of primes.

In both cases, we reach a contradiction, which means our original assumption that there are only a finite number of primes must be false. Therefore, there must be an infinite number of prime numbers.

The mystery of LLMs and structured sequences

SG Now give the proof in the style Shakespeare.

Upon a stage of numbers, prime and bold,
A tale of infinite abundance must be told.
To prove the endlessness of primes, ye see,
I summon ancient wisdom, Euclid's key.

Imagine now a list of primes so fair,
Each numbered p1, p2, and so on, in this affair.
Finite they seem, so let us now devise,
A number Q to bring a new surprise.

Q, the product of our primes, dost add one more,
For Q equals (p1 * p2 * ... * pn) + 1, and we shall explore.
Two paths unfold, a prime Q might be,
Or composite with factors we shall see.

If Q a prime, our list of primes was flawed,
For in that list, this newfound prime was not in the squad.
A contradiction lies, our assumption false,
An infinite prime existence doth exalt.

But if Q is composite, with factors in store,
We seek a prime divisor, yet unseen before.
But lo, our list of primes divides Q not well,
A remainder of one, as if under a spell.

Thus, a new prime factor we shall find,
Not in our list, a contradiction unkind.
In both our paths, our assumption laid bare,
A folly it was, for infinite primes are there.

Thus, Euclid's wisdom echoes through our play,
Infinite primes, a truth that shall not sway.

SG Now give the proof in the style of American rap.

(Verse 1)
Yo, gather 'round, let me tell you a story,
Of infinite primes and their infinite glory,
Euclid's the man, the ancient Greek master,
His prime number proof, avoidin' disaster.

Assume finite primes, a list, we create,
p1, p2, and so on, we speculate,
But hold up, let's make a number Q,
To shake things up, and find something new.

(Chorus)
Q equals the product of primes plus one,
A game changer, a riddle, we've just begun,
Infinite primes, that's the goal,
Breakin' the rules, Euclid's in control.

(Verse 2)
Now Q's either prime, or it ain't, gotta see,
Which path we take, unravel the mystery,
If Q's a prime, our list was wrong,
New prime's in the house, infinite primes strong.

But if Q ain't prime, it's gotta factor,
Prime divisors we seek, a prime-time actor,
Divide Q by our list, a remainder of one,
That ain't no good, our search ain't done.

(Chorus)
Q equals the product of primes plus one,
A game changer, a riddle, we've just begun,
Infinite primes, that's the goal,
Breakin' the rules, Euclid's in control.

Need to go back to the basics: understand mathematics of probabilistic sequences

RNNs can generate bounded hierarchical languages with optimal memory

John Hewitt, Michael Hahn, Surya Ganguli, Percy Liang, Christopher D. Manning

How Can Self-Attention Networks Recognize Dyck-n Languages?

Javid Ebrahimi, Dhruv Gelda, Wei Zhang

Unveiling Transformers with LEGO: a synthetic reasoning task

Yi Zhang, Arturs Backurs, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Tal Wagner

On the Ability and Limitations of Transformers to Recognize Formal Languages

Satwik Bhattamishra♠ Kabir Ahuja♢* Navin Goyal♠
♠ Microsoft Research India

Thinking Like Transformers

Gail Weiss [1] Yoav Goldberg [2 3] Eran Yahav [1]

Teaching Arithmetic to Small Transformers

Nayoung Lee, Kartik Sreenivasan, Jason D. Lee, Kangwook Lee, Dimitris Papailiopoulos

Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression

Allan Raventós, Mansheej Paul, Feng Chen, Surya Ganguli

Outline: today's frontiers and tomorrow challenges/opportunities

(1) Alternate physical implementations of ML algorithms: negotiating high dimensional error landscapes with atoms and photons in open dissipative open systems.

(2) The physical origins of generative AI: from diffusion models to societal implications.

(3) Mechanistic interpretability: a unification of neuroscience, physics and AI with applications to alignment and government regulation

(4) Theory: the generalization puzzle: how can huge networks generalize without overfitting.

(5) Unsustainable neural scaling laws and the combatting the data hungriness of modern AI

(6) The mystery of structured sequences: the unreasonable effectiveness of predicting every next word on the internet (+ human feedback). LLMs as the next challenge/opportunity for theory and society.