ML, QML, and Dynamics

What mathematics can help us understand and advance machine learning?

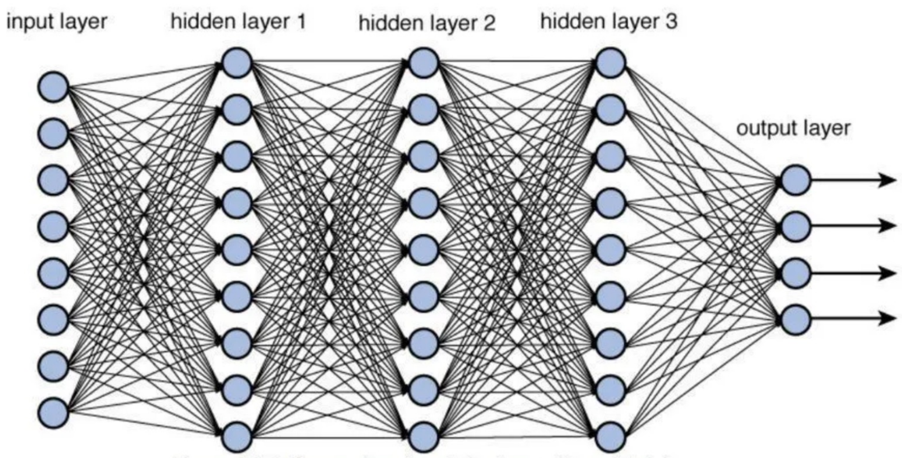
Michael Freedman

CMSA Harvard

October 28 2023

DNN

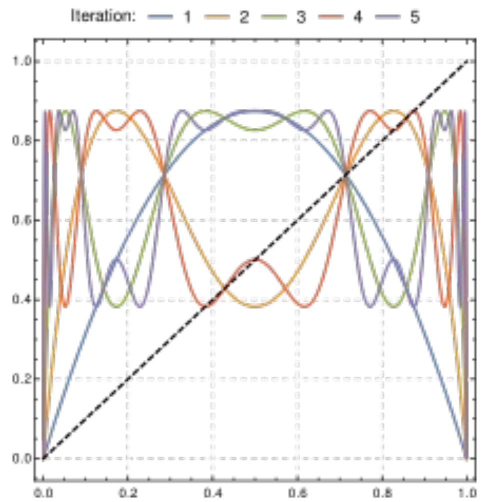
- Vanilla Deep Neural Nets (DNNs) are reminiscent of the stretch and fold dynamics that lead to chaos.

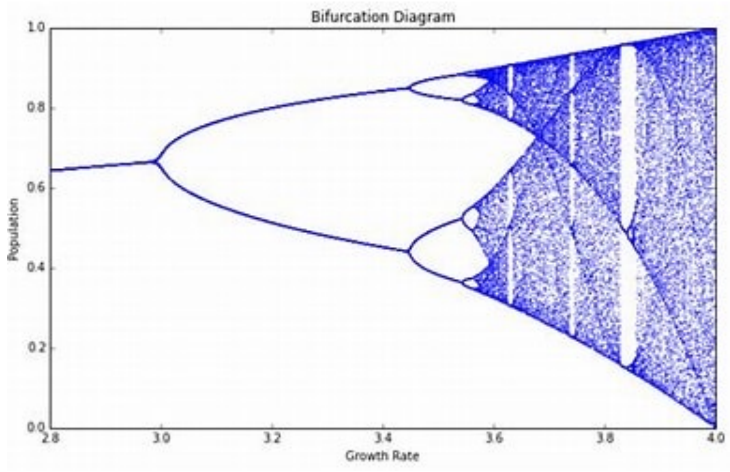


- Each step is a linear map L followed by σ, a coordinate-wise non-linear activation function: $(\sigma \circ Lin_D) \circ \cdots \circ (\sigma \circ Lin_1)$.
- These linear maps typically have singular values both >1 and <1.

Logistic Map

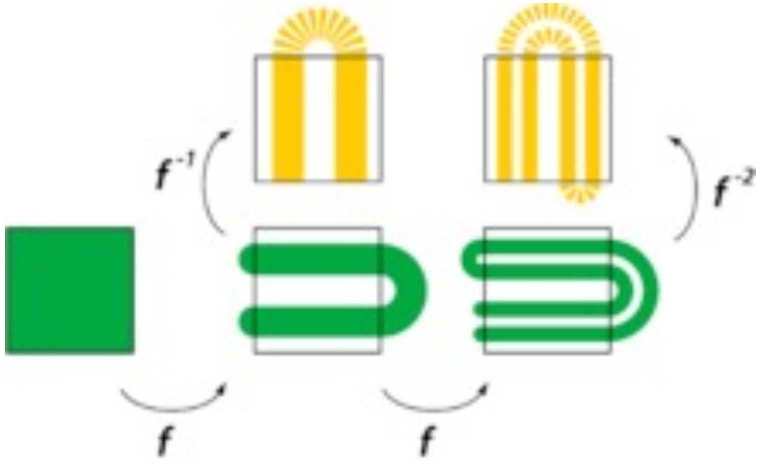
- $x_{n+1} = -r\left(x_n(x_n - 1)\right), r > 0$
- Feigenbaum (4.6…): period doubling, then chaos
- Transition to chaos
- Pomeau-Manville scenario
- Intermittency
- York-Li period 3 ⇒ chaos





Smale Horseshoe Map

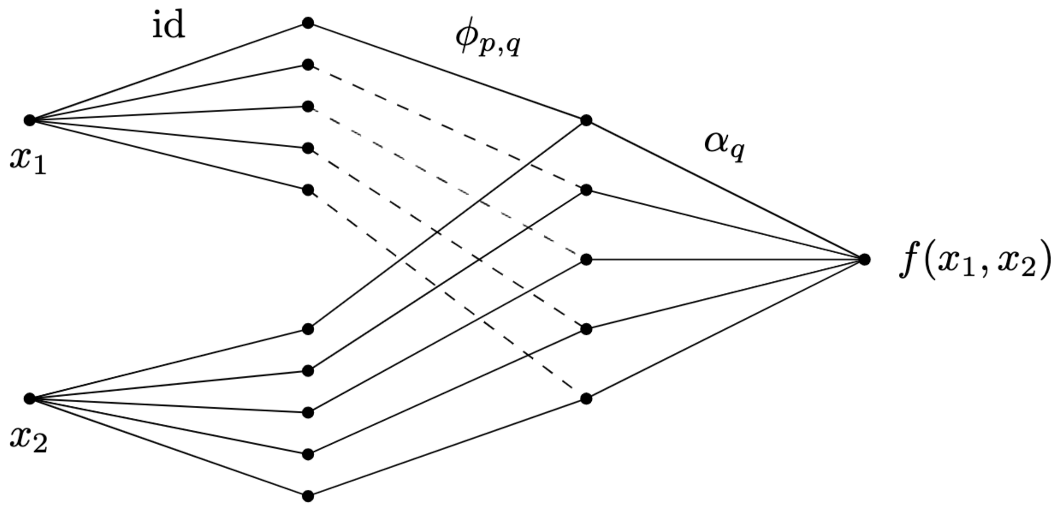
- High-order yellow and green stripes intersect in an invariant Cantor set with chaotic “hyperbolic” dynamics.
- A foundational example of symbolic dynamics



Expressibility of Functions (Hilbert's 13th Problem)

Kolmogrov-Arnold representation theorem (1957)

Given $n \geq 2$, $\exists n(2n+1)$ continuous functions $\phi_{p,q}: I \to I$, $1 \leq p \leq n$, $1 \leq q \leq 2n+1$ such that for continuous $f: I^n \to I$, there exists continuous $\alpha_q: I \to I$, $1 \leq q \leq 2n+1$ s.t. $f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} \alpha_q\left(\sum_{p=1}^{n} \phi_{p,q}(x_p)\right)$ (the $\phi_{p,q}$ can be chosen Lipschitz).



Note. NN has depth $D = 3$ but extravagant nonlinearities. A trade-off in expressibility and trainabilty.

Expressibility of Functions

Sternberg-Ostrand (1989)

For $n > 2$, X is a compact metric space of dim $\leq n$ iff $\exists$ an embedding $X \hookrightarrow \mathbb{R}^{2n+1}$, $x \mapsto y$, which is “basic,” meaning all continuous $g\colon X \to I$ can be written

$$g_{(x)} = \sum_{q=1}^{2n+1} \alpha_q(y_q)$$

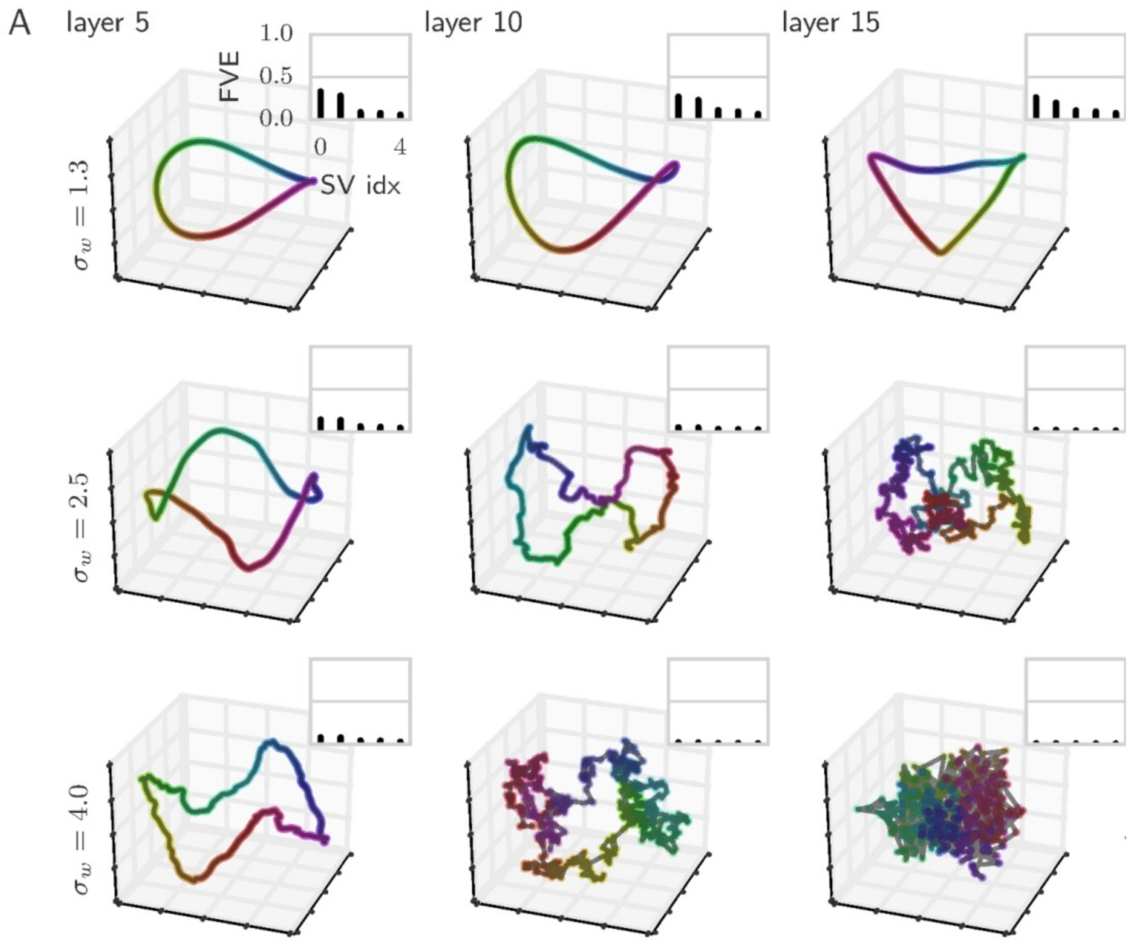
For some continuous $\{\alpha_q\colon I \to I\}, 1 \leq q \leq 2n+1.$

Expressibility of Functions

- But AI *has mostly ignored this subject* to date. The α_q and $\phi_{p,q}$ have been regarded as too pathological to train.
- I would like to see a quantitative version of Kolmogorov-Arnold-Ostrand-Sternberg theory.

Expressivity in DNN

- Poole, Lahiri, Raghu, Sohl-Dickstein, Ganguli (2016) study the evolving differential geometry of Gaussian random DNN.
- They report chaotic evolution when ${}^{\sigma_w}/_{\sigma_b} \gg 1$, where σ_w = weight variance, σ_b = off-set variance.



Expressivity in DNN

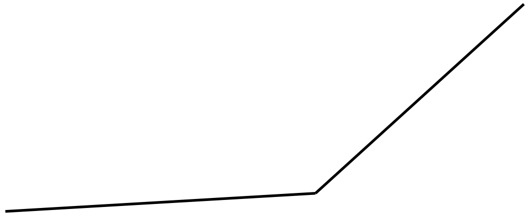
Question: Are these chaotic graphics approaching a basic embedding?
Can such an approach be defined in terms of expressivity? Trainability?

Laps Conjecture

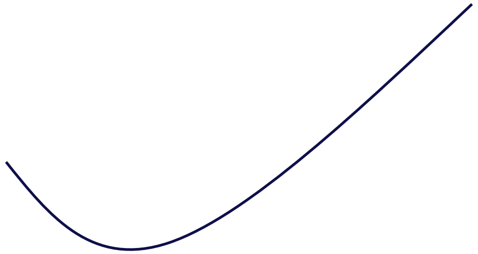
Laps Conjecture (very rough)

Expressivity :: $\text{depth}(\log(\text{laps}+1)) \approx \log(\#\ \text{pastry layers})$

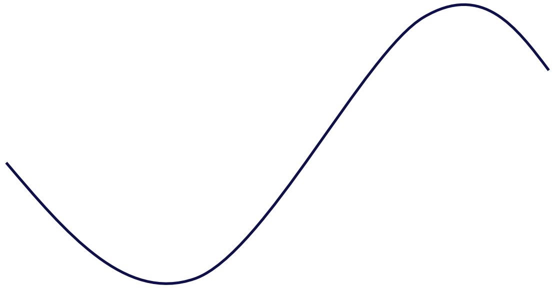
# of pastry layers = $\text{laps}^{\text{depth}}$



1 lap



2 laps



3 laps

Power of Data in Quantum Machine Learning

- Huang, Broughton, Mohseni, Babbush, Boixo, Neven, McClean (2022): feature (or token) embedding space

Classical

- Naively $\mathbb{R}^{1000}$
- Lots of nonlinear dynamics prior to linear separation

Quantum

- $\mathbb{C}^{(2^{1000})}$
- Each picture a different basis state
- Linear separation

Caveats

1. Reproducing kernel methods effectively expand classical embedding space. $\langle f, g\rangle_k = \iint dx\, dy\, f(x)K(x, y)g(y)$
2. **Lindenstrauss Theorem.** $n \approx c\frac{\log N}{\epsilon^2}$ is sufficient dimension to embed a finite metric space X of n points up to inaccuracy ϵ.

What Could Quantum Offer ML

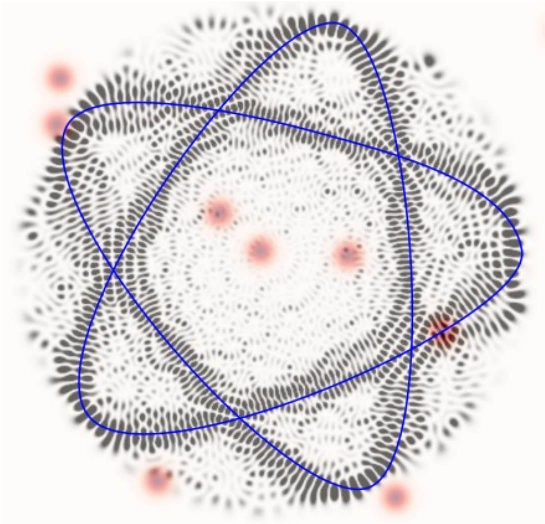
- Work of S. lloyd and others have shown super-polynomial qantum speed-ups in certain linear algebra tasks.
- This may be useful in the longer term, but is there anything noisy, near-term quantum computers might offer?

What Could Quantum Offer ML

- Assuming that chaos is key to ML, it may help to have two separate separate understandings.
- I’ve mentioned the classical picture: orbit based.
- There is the spectral picture of quantum mechanics, for which early quantum computers will be an experimental window.

What Could Quantum Offer ML

- From quantum scar—Wikipedia: quantum scars along an unstable classic orbit.
- Guttwiller (~Selberg) trace formula:



$$R = tr\left(\frac{1}{H-E}\right), R_{sc}(E) = \sum_{k\ closed\ orbit} T_k \sum_{n=1}^{\infty} \frac{1}{2\sinh(\chi_{n,k})} e^{i\left(nS_k(E) - \alpha_{n,k}\frac{\pi}{2}\right)}$$

semi-classical resolvant

length of orbit

density of orbit

action along k[th] closed orbit at energy E

Maslow index

What Could Quantum Offer ML

- Perhaps we can learn what kinds of chaos are useful for learning. If so, noisy quantum computers/quantum simulators offer and orthogonal approach to creating such dynamics.
- This is the subject of “active matter”: Floquette phases, measurement induced phase transitions, streamlining the measurement-action loop.
- In the limit of precision and control, active matter becomes a quantum computer.

More Math for LLMs

- What should mathematicians look for in the functioning of large language models (LLMs)?
- I'll offer two suggestions:

1. Emergent tensor structures in trained linear maps: W, Q, K, V, etc.
2. Orders *not* induced by a potential function

More Math for LLMs

- With collaborators Moj Shokrian-Zini and Adam Brown, we found when minimizing geometric loss functions such as *Ricci scalar curvature* over left invariant metrics on $\mathrm{SU}(2^n)$ that the principal axis often respected an emergent tensor structure.
- For example, minimizing $R(g_{ij})$ on $\mathrm{SU}(8)$, all 63 principal axes $\{iH_k, 1 \leq k \leq 63\}$ take the form $H_k \approx j(H_{1,k} \otimes H_{2,k} \otimes H_{3,k})$ for $1 \leq k \leq 63$, $H_{i,k} \in$ 2×2 Hermitian h_2, and $j\colon (h_2 \otimes h_2 \otimes h_2) \cong h_8$ induced by some fixed, emergent $J\colon \mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \mathbb{C}^2 \to \mathbb{C}^8$.

See: Michael Freedman & Mojtaba Shokrian-Zini, “The universe from a single particle I-III,” *Journal of High Energy Physics* (2021) and Mojtaba Shokrian-Zini, Adam Brown, & Michael Freedman, “The smallest interacting universe,” *Journal of High Energy Physics* (2023).

More Math for LLMs

- Like this example from high energy, one may look for emergent tensor structures within the highly trained linear operators between inner product spaces, e.g., vanilla weight matrices W, or the specialized transformer ingredients: Query, Key, and Value (Q, K, V)

More Math for LLMs

- Given a linear map L between inner product spaces, $L: A \to B$, we may ask among all isomorphisms $A' \otimes A'' \overset{\cong}{\to}_a A$, $B' \otimes B'' \overset{\cong}{\to}_b B$, to minimize the tensor rank of $b^{-1} \circ L \circ a \in \mathrm{Hom}(A' \otimes A'', B' \otimes B'')$ across the middle $\otimes$ in $b^{-1} \circ L \circ a \in \left(A'^{*} \otimes B'\right) \otimes (A''^{*} \otimes B'')$.
- Notice this is quite distinct from the usual notion of matrix rank, which is rank across the middle tensor in $(A' \otimes A'')^{*} \otimes (B' \otimes B'')$.

More Math for LLMs

- Curiously, "tensor rank"$= 1$ corresponds to the solution of a highly over-determined linear system at the heart of the *Kolmogorov-Arnold representation theorem*.
- When is an $n \times m$ array of real numbers the sum of values written along the axes? For example, for $n = m = 2$:

	a	b
c	1	-1
d	-1	1

cannot be, so $\{\log(\text{singular values}) = 1,1,-1,-1\}$ is *incompatible* with tensor rank 1.

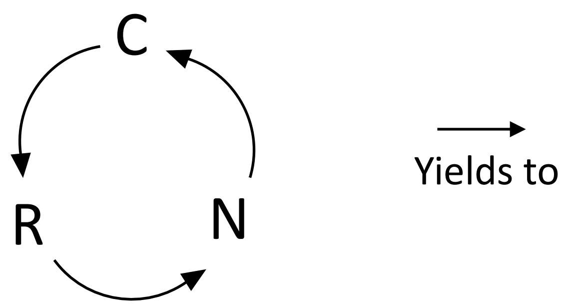
More Math for LLMs

- I’m currently doing numerics of the singular values of ML matrices with a U. of Maryland group: Barkeshli, He, and Kalra, looking for tensor structures which stand out above the noise background.
- A French school in cognitive science (Slomensky et al.) has proposed iterated tensor structures in language, but in a fixed basis, i.e., nouns and verbs, “roles” and “fillers” rather than an emergent one.

Non-gradient-like Orders

Rock Paper Scissors (RPS)

Ecology: E. coli {
- colicin producing
- resistant
- naive

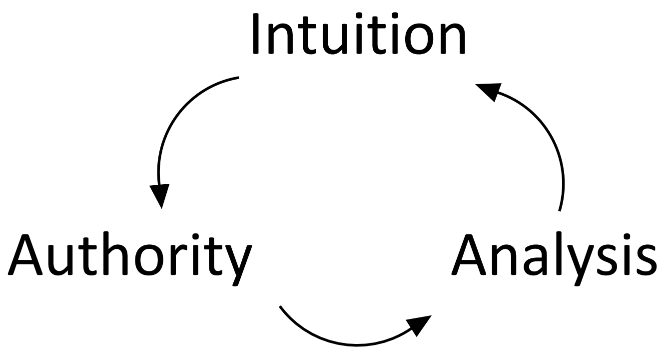


RPS Cellular Automata

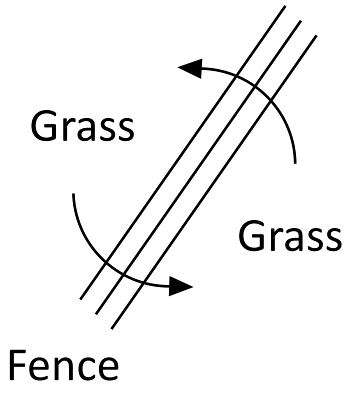
RPS Cellular Automata

- “A million species in every shovelful of dirt” ⇒ competition beyond potential functions.

Ecology of mind:



2-cycles:



Kahnemann-Tversky, Prospect Theory (1979)
Notable reply: Mercier-Sperber, *Enigma of Reason*

Conclusion

- Classical ML seems to have little use for $i = \sqrt{-1}$.
- Perhaps QML is better suited to generate cyclic orders like e^{it}, where chiral models are readily available.
- Look for emergent structures within the linear algebra
- Understand how dynamics can inform ML
- Quantify Kolmogorov–Arnold theory (Laps conjecture).