Mathematical and sociological questions in LLM Theory, Deep Learning Theory.

Matus Telgarsky, Courant Institute, NYU.

Thanks to: ■Sivaraman Balakrishnan (CMU), ■Misha Belkin (UCSD), ■Margalit Glasgow (MIT), ■Daniel Hsu (Columbia), ■Audrey Huang (UIUC / MSR NYC), ■Ziwei Ji (Google Research), ■Akshay Krishnamurthy (MSR NYC), ■Lana Lazebnik (UIUC), ■Maxim Raginsky (UIUC), ■Jingfeng Wu (UC Berkeley), ■Fanny Yang (ETH Zürich), ■Fall 2024 Simons Institute ML Program Participants, ■o3, ■`o4-mini-high`, ■`Opus`.

Plan for today.

- **Technical questions.**
- Speculative questions.
- Sociological questions.

Scope: no safety, (length) generalization, interpretability, ...

Open problem #1: GD learns features.

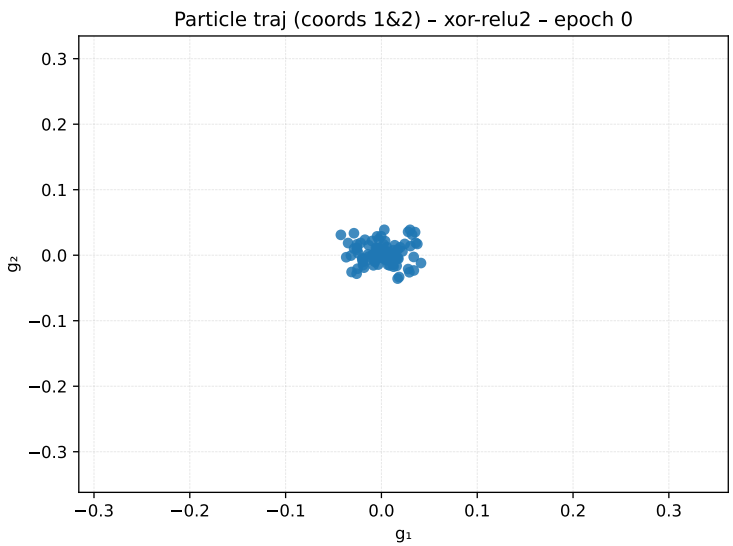
Necessary and sufficient conditions on feature learning.

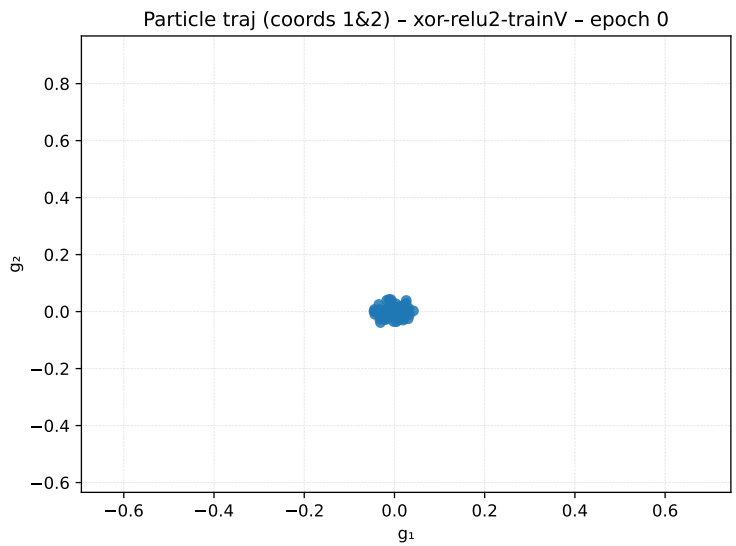
Why we care:

- (Math) Interesting analytic tools are missing.
- (Empirical) Original promise of deep learning.

2-XOR problem (Wei-Lee-Liu-Ma, 2018).

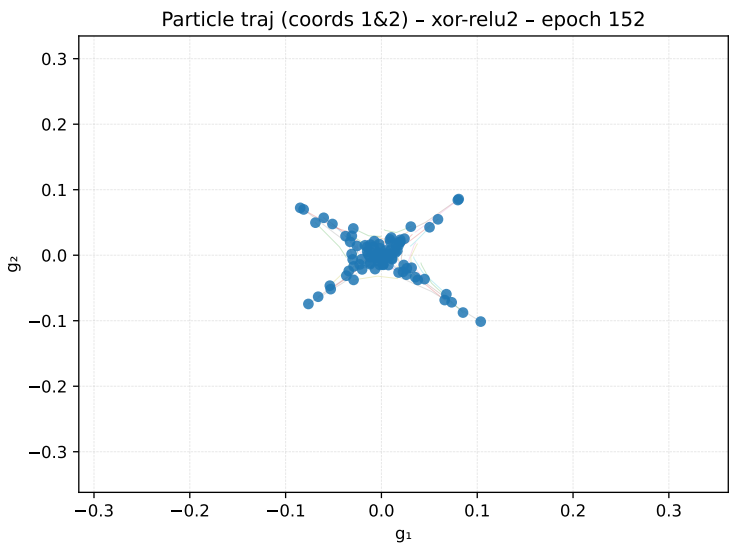
- 2-ReLU network: $x \mapsto \sum_{j=1}^{m} a_j \max\{0, v_j^\top x\}$.
- $x \sim \text{Uniform}\left((\pm\frac{1}{\sqrt{d}})^d\right)$.
- $y = d \cdot x_1 \cdot x_2$.
- Logistic loss $z \mapsto \ln(1 + e^{-z})$.
- "Particles": $\forall j, \quad t \mapsto a_j(t)v_j(t)$.

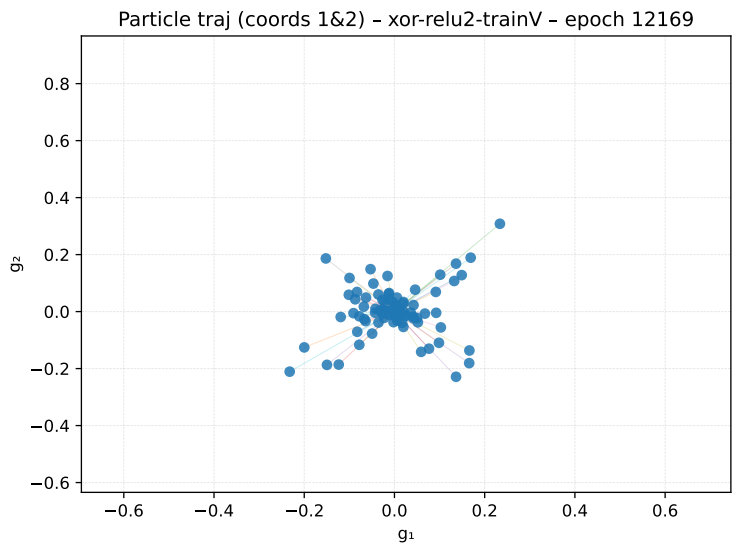




2-XOR problem (Wei-Lee-Liu-Ma, 2018).

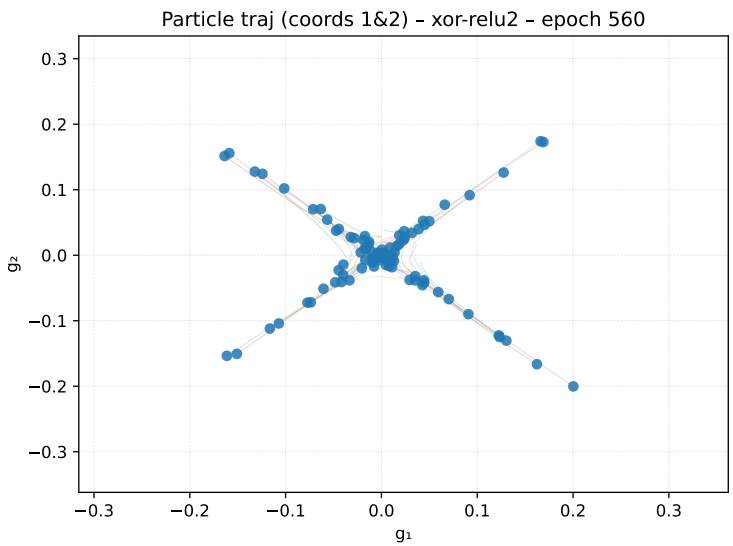
- 2-ReLU network: $x \mapsto \sum_{j=1}^{m} a_j \max\{0, v_j^\top x\}$.
- $x \sim \text{Uniform}\left((\pm\frac{1}{\sqrt{d}})^d\right)$.
- $y = d \cdot x_1 \cdot x_2$.
- Logistic loss $z \mapsto \ln(1 + e^{-z})$.
- "Particles": $\forall j, \quad t \mapsto a_j(t)v_j(t)$.

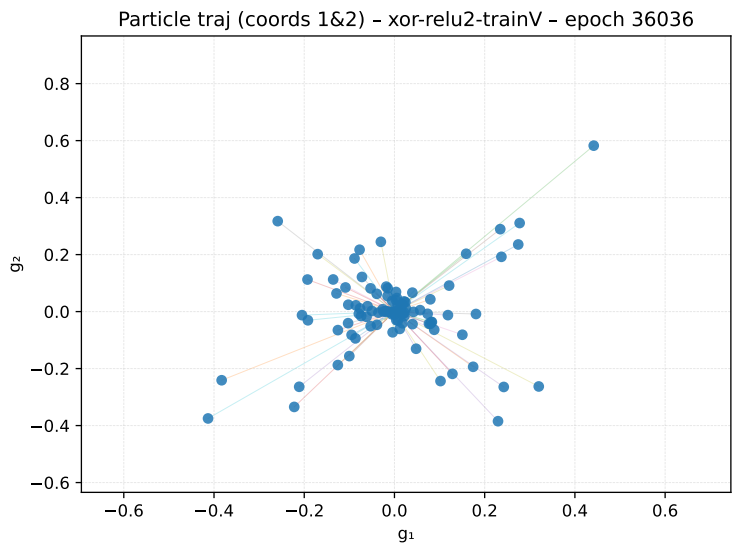




2-XOR problem (Wei-Lee-Liu-Ma, 2018).

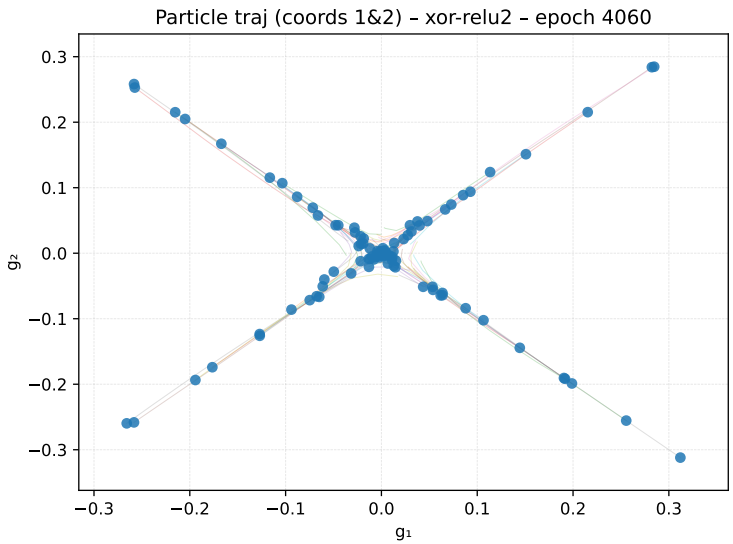
- 2-ReLU network: $x \mapsto \sum_{j=1}^{m} a_j \max\{0, v_j^\top x\}$.
- $x \sim \text{Uniform}\left((\pm\frac{1}{\sqrt{d}})^d\right)$.
- $y = d \cdot x_1 \cdot x_2$.
- Logistic loss $z \mapsto \ln(1 + e^{-z})$.
- "Particles": $\forall j, \quad t \mapsto a_j(t)v_j(t)$.

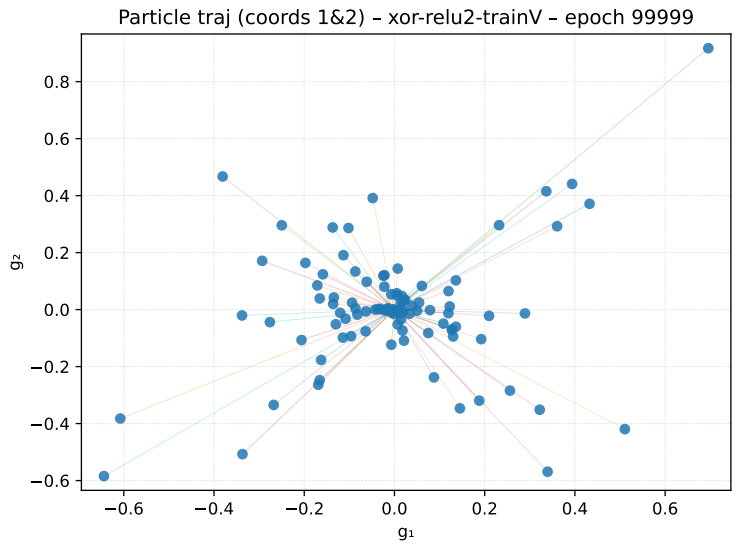


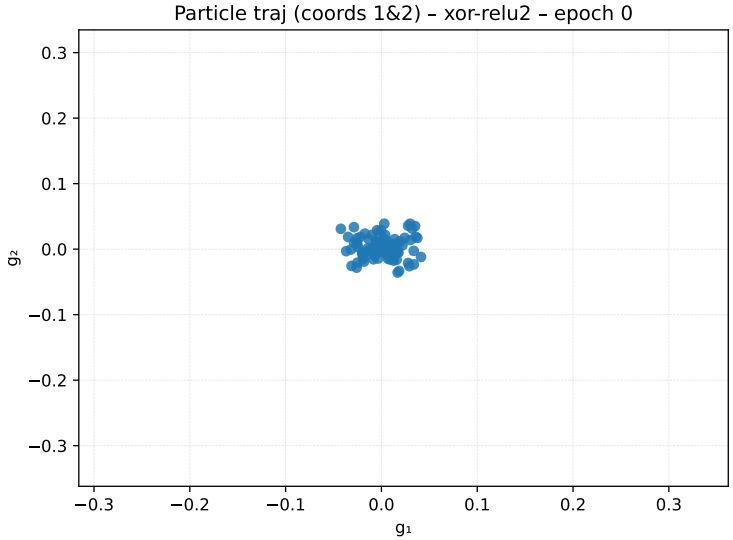


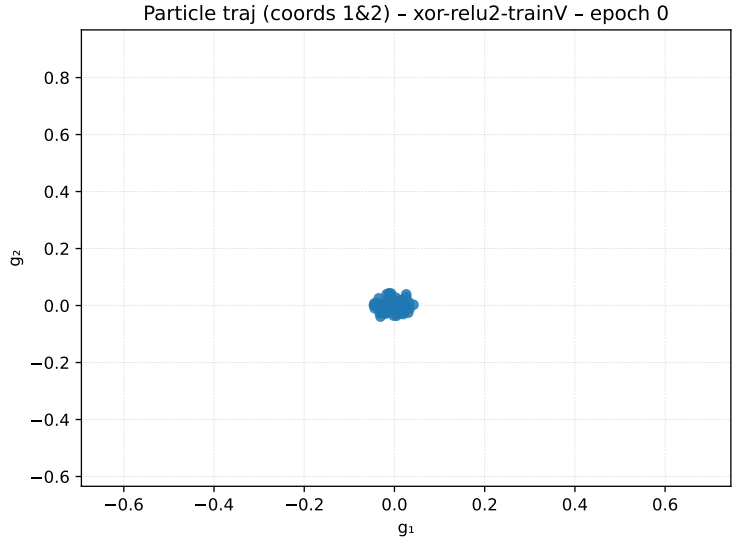
2-XOR problem (Wei-Lee-Liu-Ma, 2018).

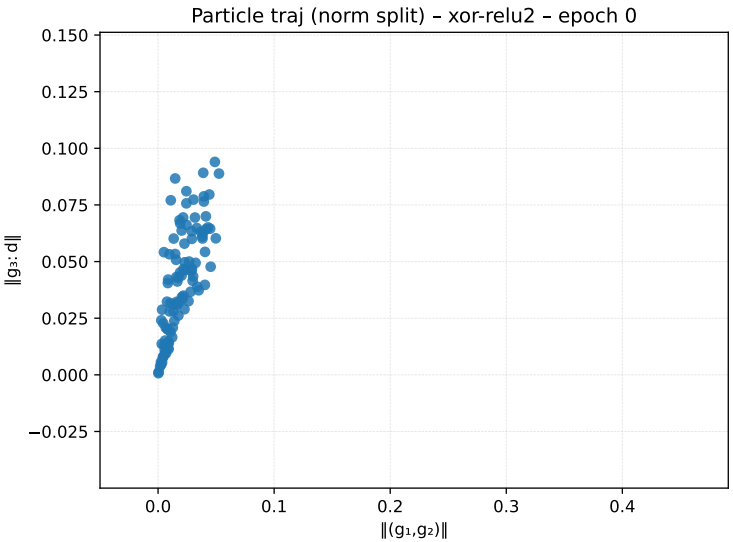
- 2-ReLU network: $x \mapsto \sum_{j=1}^{m} a_j \max\{0, v_j^\top x\}$.
- $x \sim \text{Uniform}\left((\pm\frac{1}{\sqrt{d}})^d\right)$.
- $y = d \cdot x_1 \cdot x_2$.
- Logistic loss $z \mapsto \ln(1 + e^{-z})$.
- "Particles": $\forall j, \quad t \mapsto a_j(t)v_j(t)$.

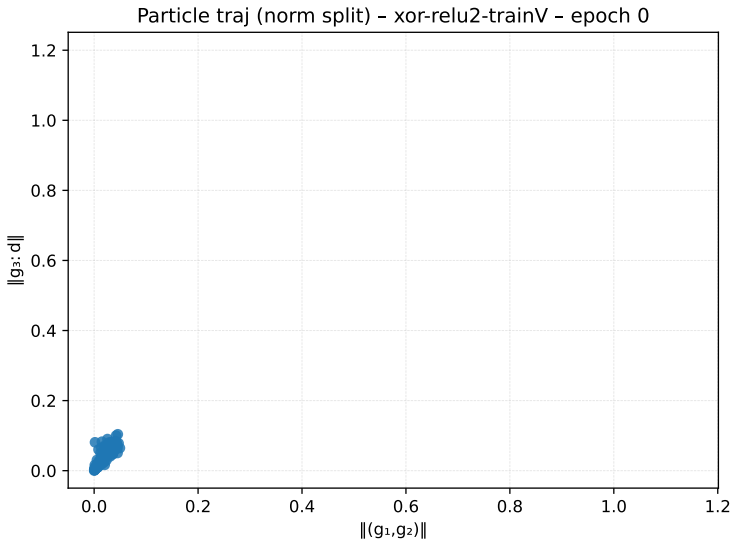


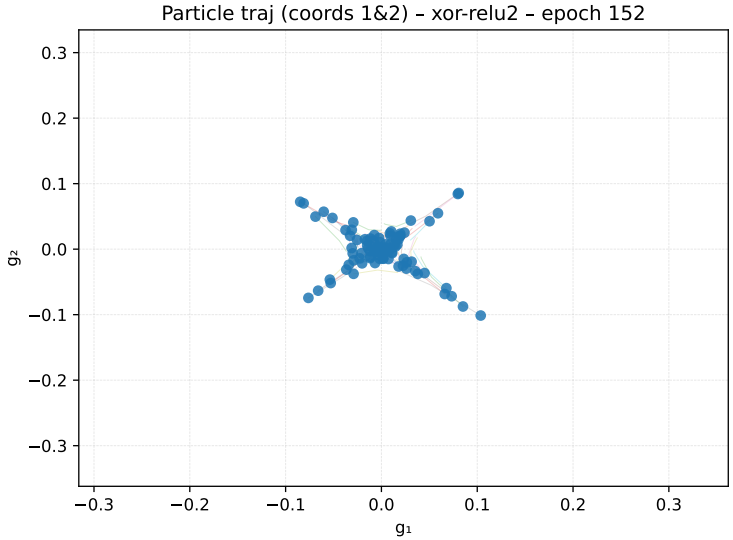


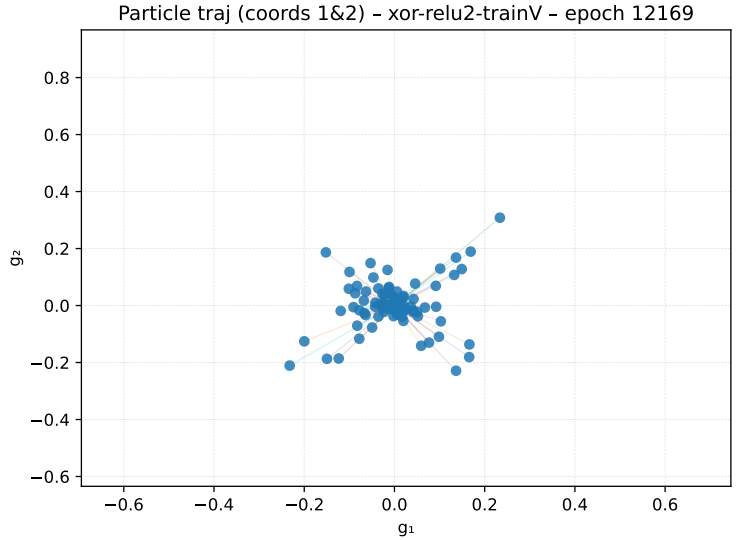

Particle traj (coords 1&2) – xor-relu2 – epoch 0
g₂
g₁

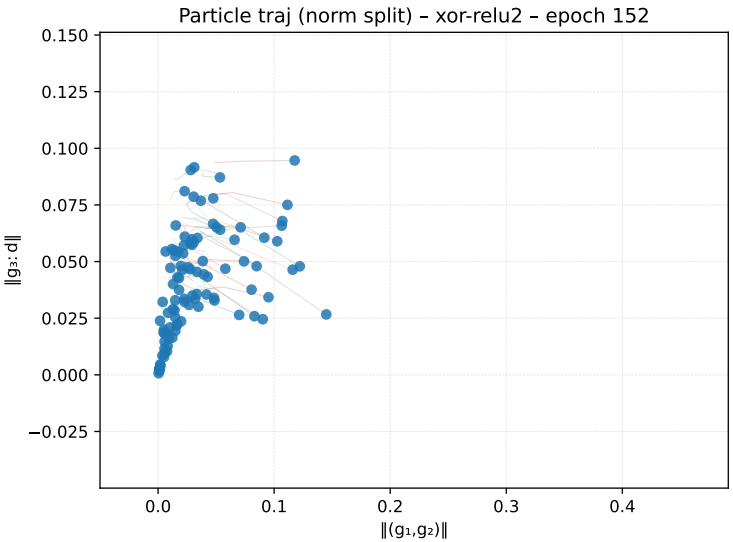

Particle traj (coords 1&2) – xor-relu2-trainV – epoch 0
g₂
g₁

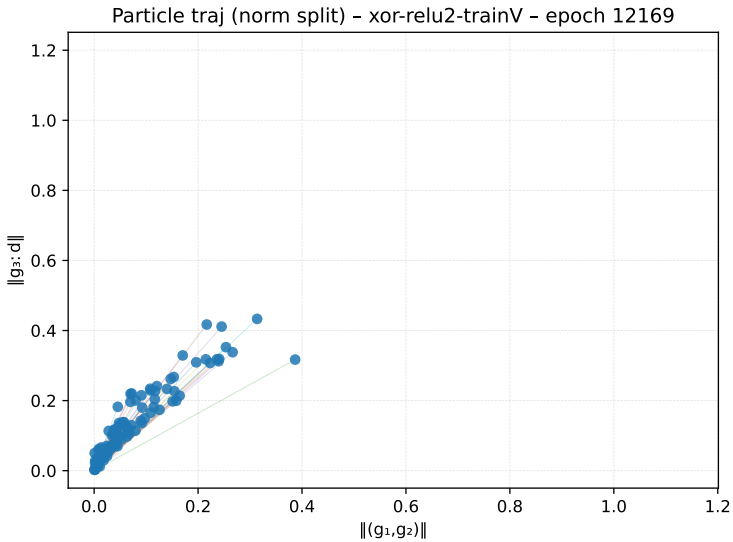

Particle traj (norm split) – xor-relu2 – epoch 0
‖g₃:d‖
‖(g₁,g₂)‖

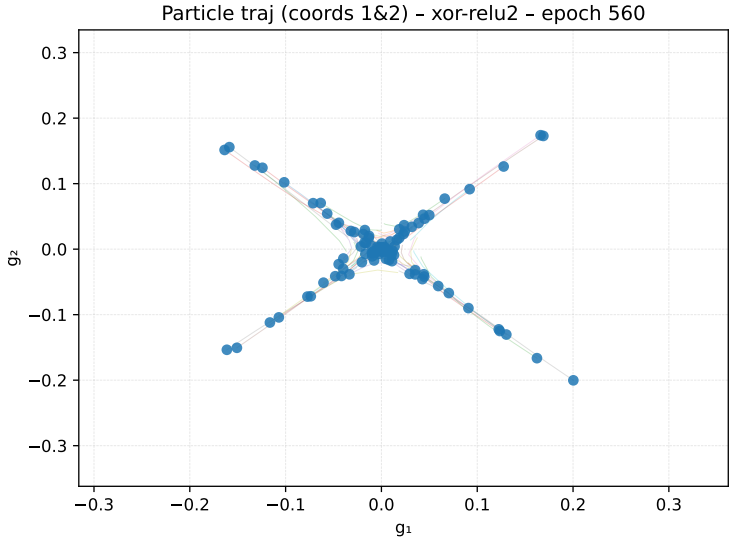

Particle traj (norm split) – xor-relu2-trainV – epoch 0
‖g₃:d‖
‖(g₁,g₂)‖

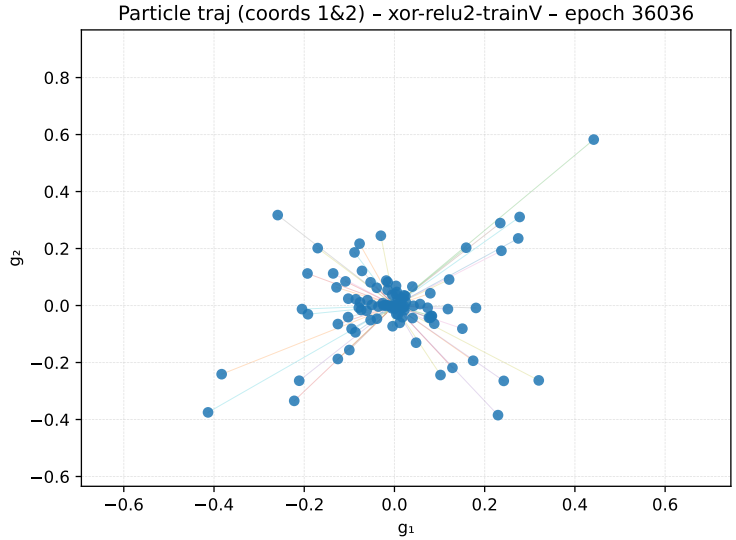

Particle traj (coords 1&2) – xor-relu2 – epoch 152
g₂
g₁

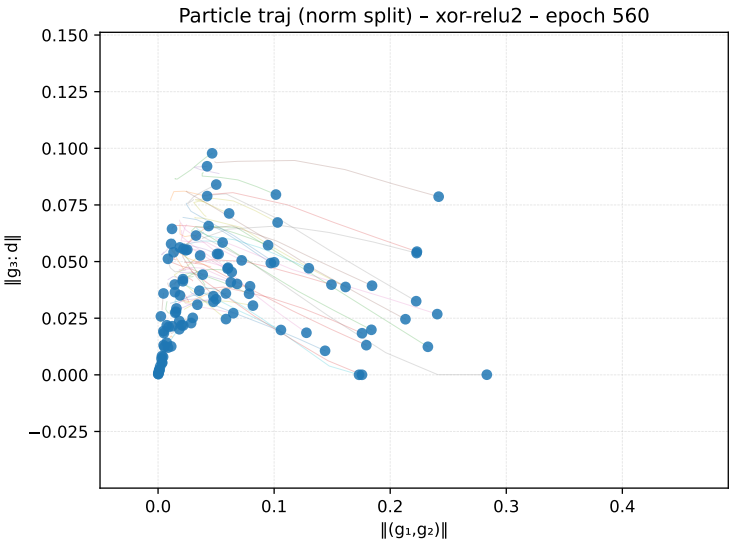

Particle traj (coords 1&2) – xor-relu2-trainV – epoch 12169
g₂
g₁

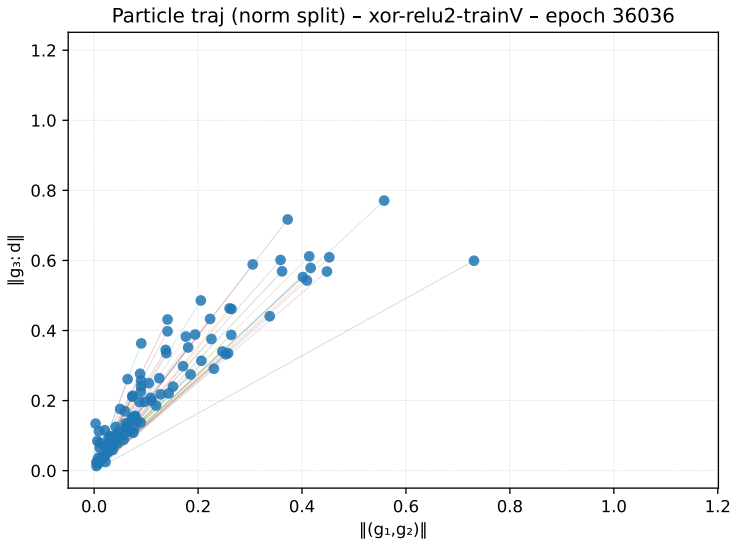

Particle traj (norm split) – xor-relu2 – epoch 152
‖g₃:d‖
‖(g₁,g₂)‖

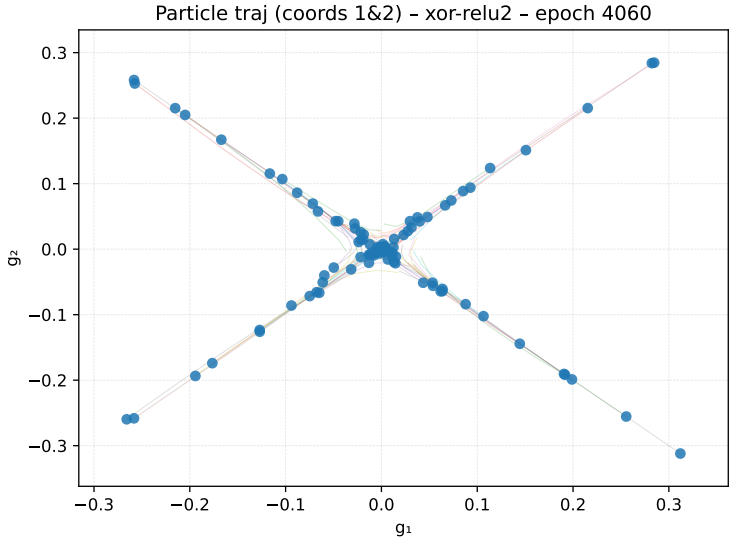

Particle traj (norm split) – xor-relu2-trainV – epoch 12169
‖g₃:d‖
‖(g₁,g₂)‖

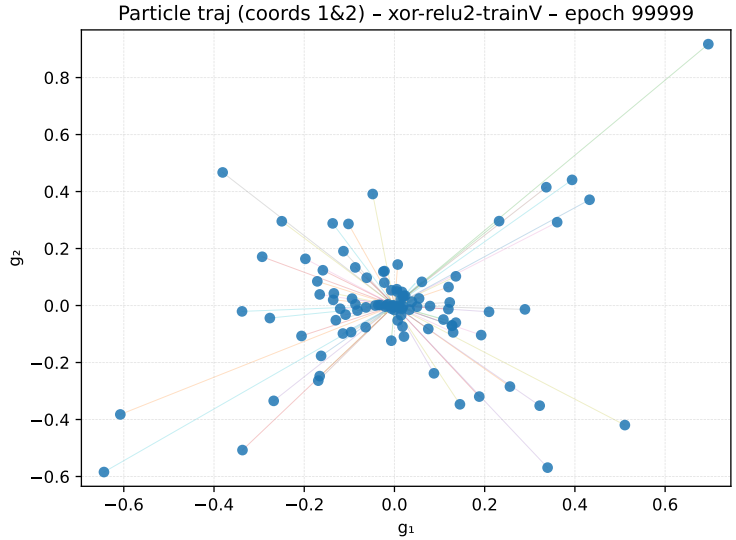

Particle traj (coords 1&2) – xor-relu2 – epoch 560
g₂
g₁

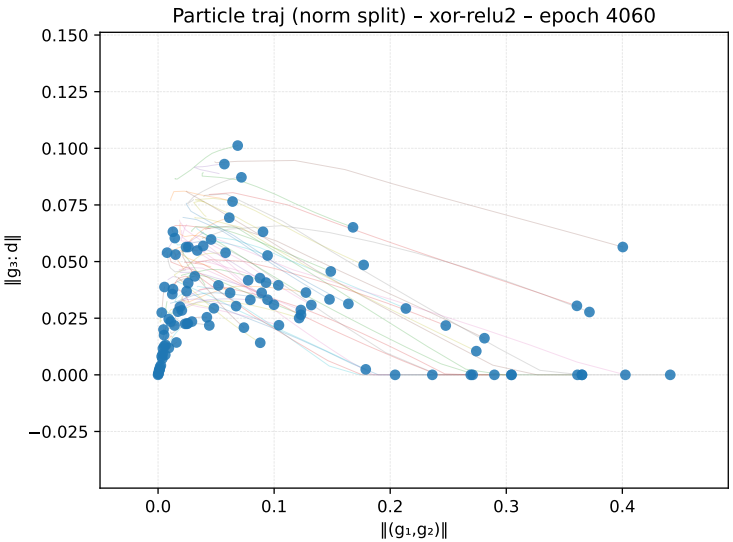

Particle traj (coords 1&2) – xor-relu2-trainV – epoch 36036
g₂
g₁

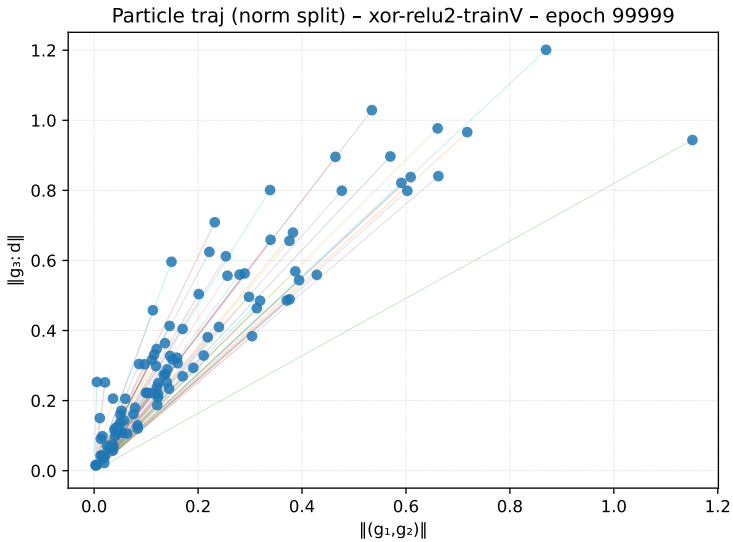

Particle traj (norm split) – xor-relu2 – epoch 560
‖g₃:d‖
‖(g₁,g₂)‖


Particle traj (norm split) – xor-relu2-trainV – epoch 36036
‖g₃:d‖
‖(g₁,g₂)‖


Particle traj (coords 1&2) – xor-relu2 – epoch 4060
g₂
g₁


Particle traj (coords 1&2) – xor-relu2-trainV – epoch 99999
g₂
g₁


Particle traj (norm split) – xor-relu2 – epoch 4060
‖g₃:d‖
‖(g₁,g₂)‖


Particle traj (norm split) – xor-relu2-trainV – epoch 99999
‖g₃:d‖
‖(g₁,g₂)‖

Open ("forgotten") problem #1a: general NTK separation.

Necessary and sufficient conditions on feature learning.
a. Separate #samples for $\mathcal{L}(f_t)$ vs $\mathcal{L}(f_0 + \langle \nabla f_0, w_t - w_0 \rangle)$.

Known:

- Only general tool (NTK)
  is our no-feature-learning baseline.
- 2-XOR + 2-layer ReLU + SGD (Margalit Glasgow, 2024);
  breaks with standard init or (≥ 3)-XOR.
- Mean-field: special data + 2-layer + modified GD.
- Margins: conditional on init, approximate homogeneity
  (Srebro-. . . , Lyu-Li, Ji-Telgarsky, Cai-W-M-L-B).

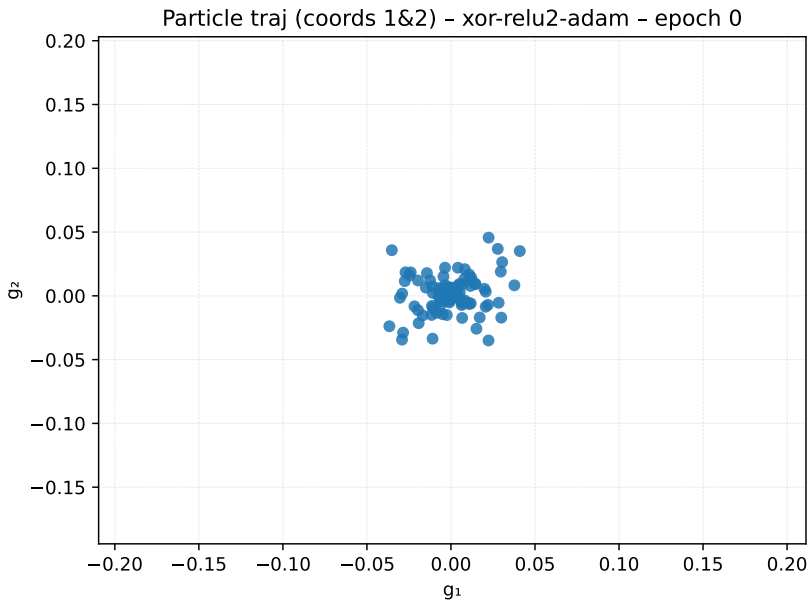
Unknown:

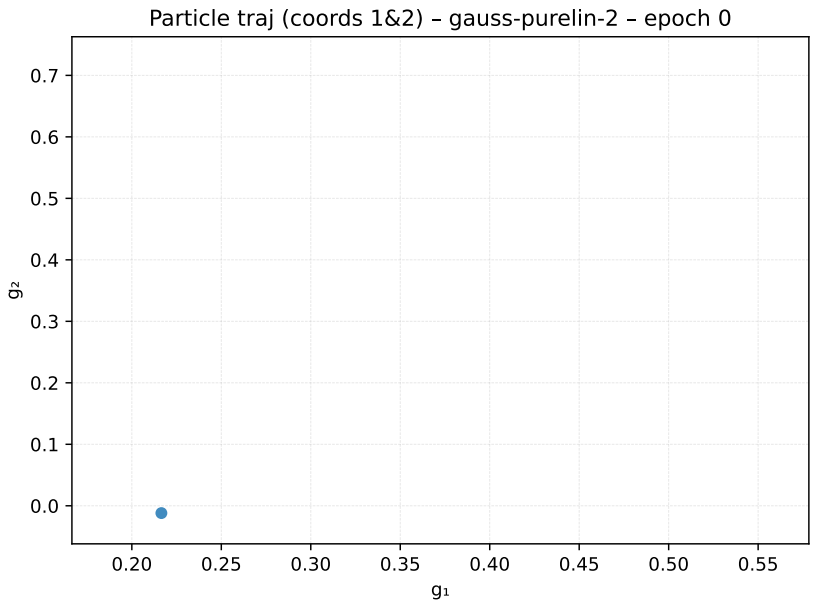
- Practical shallow/deep MLP, transformer, . . .

Path following.

- 2-XOR follows a clean path.
- Linear regression follows a cleaner path:
  given data $X = \sum_i s_i u_i v_i^\mathsf{T}$ and optimal w,

$$v_i^\mathsf{T} (w_t - w) = (1 - \eta s_i^2)^t v_i^\mathsf{T} (w_0 - w) .$$

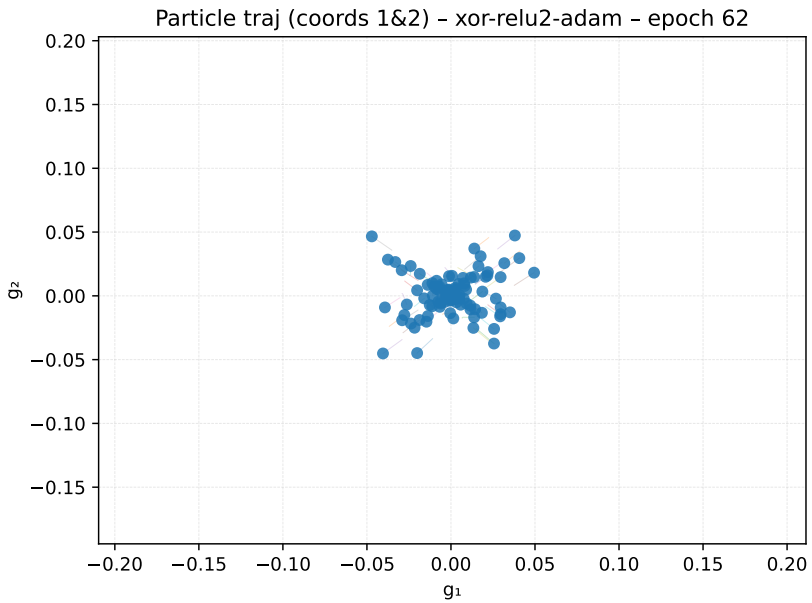


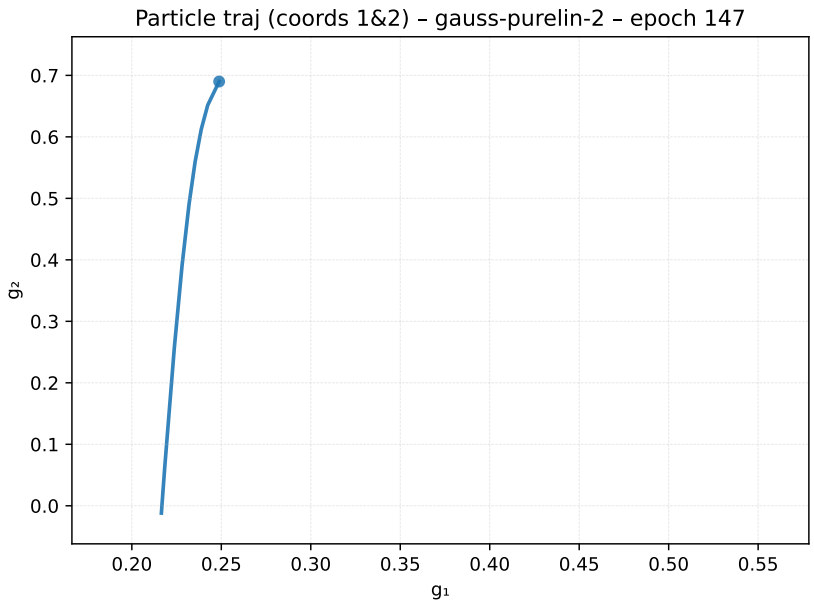


Path following.

- 2-XOR follows a clean path.
- Linear regression follows a cleaner path:
  given data $X = \sum_i s_i u_i v_i^\mathsf{T}$ and optimal w,

$$v_i^\mathsf{T}(w_t - w) = (1 - \eta s_i^2)^t v_i^\mathsf{T}(w_0 - w).$$

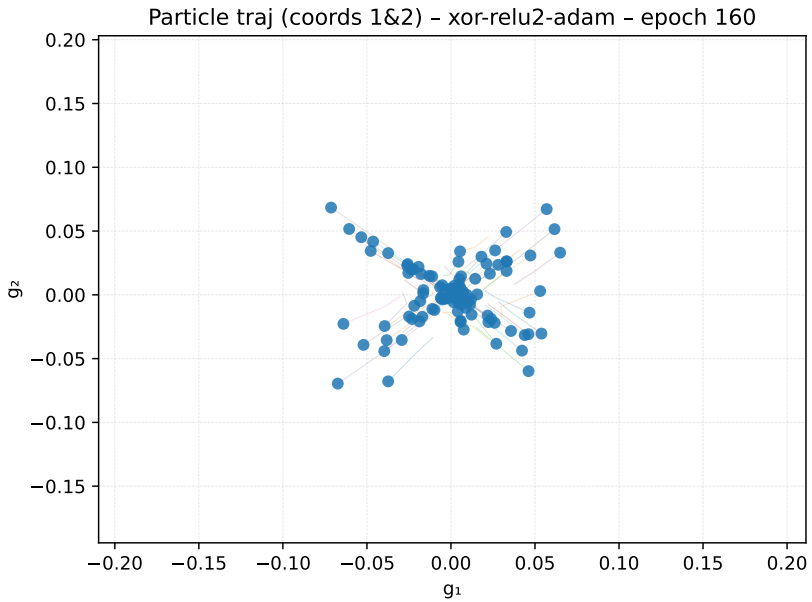


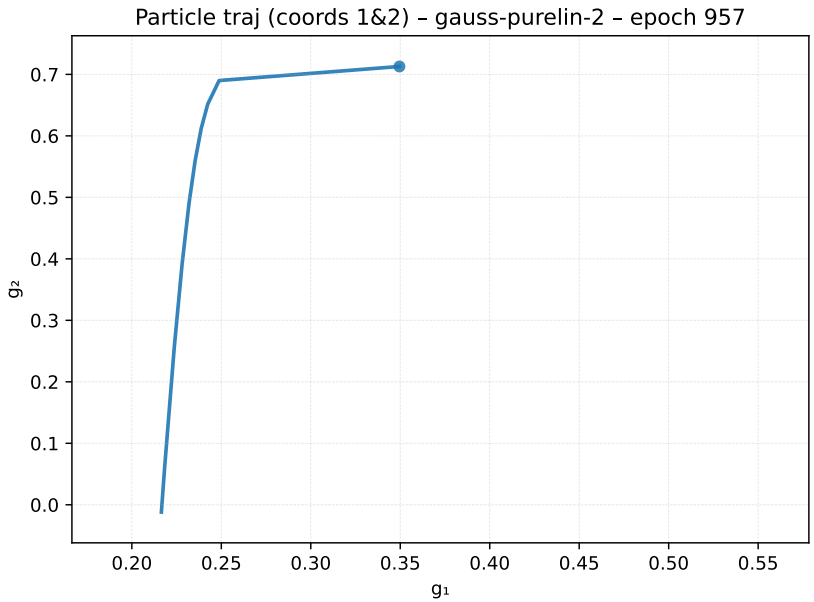


Path following.

- 2-XOR follows a clean path.
- Linear regression follows a cleaner path:
  given data $X = \sum_i s_i u_i v_i^\mathsf{T}$ and optimal w,

$$v_i^\mathsf{T}(w_t - w) = (1 - \eta s_i^2)^t v_i^\mathsf{T}(w_0 - w).$$

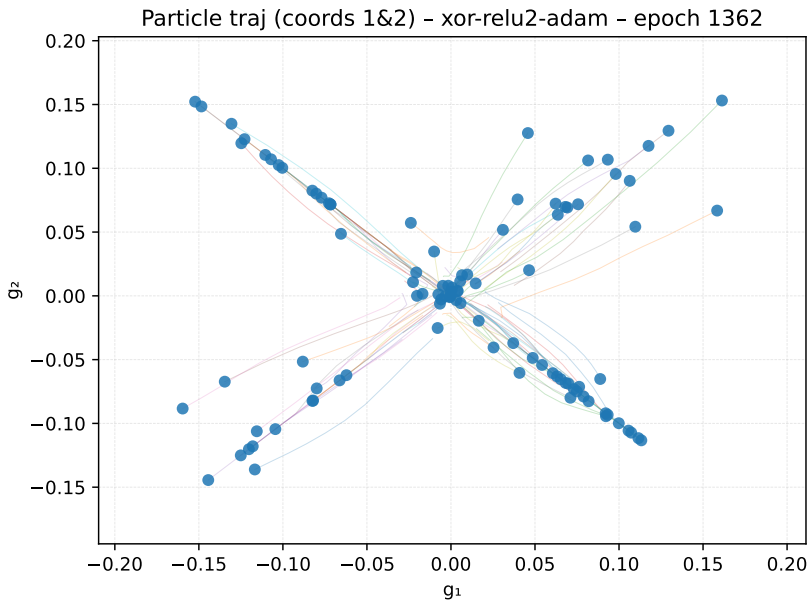


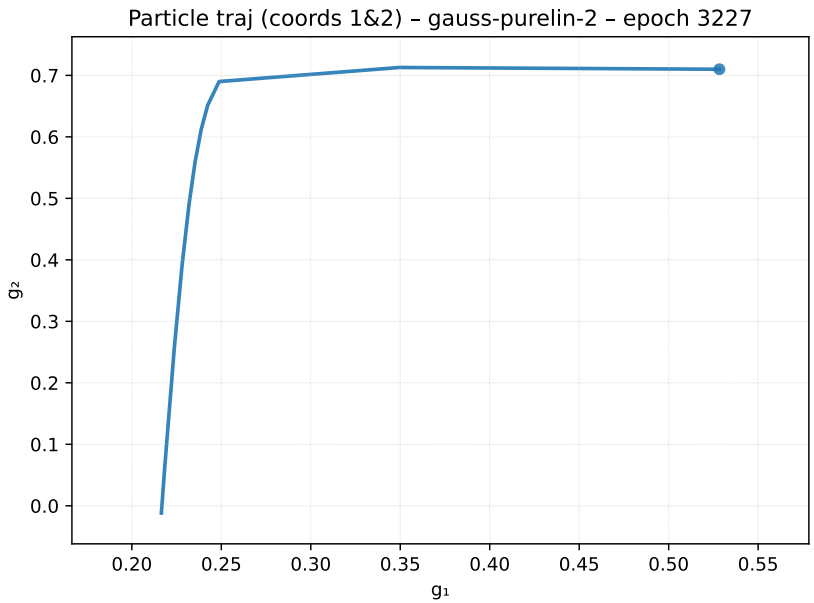


Path following.

- 2-XOR follows a clean path.
- Linear regression follows a cleaner path:
  given data $X = \sum_i s_i u_i v_i^\mathsf{T}$ and optimal w,

$$v_i^\mathsf{T}(w_t - w) = (1 - \eta s_i^2)^t v_i^\mathsf{T}(w_0 - w).$$





Open problem #1b: path following.

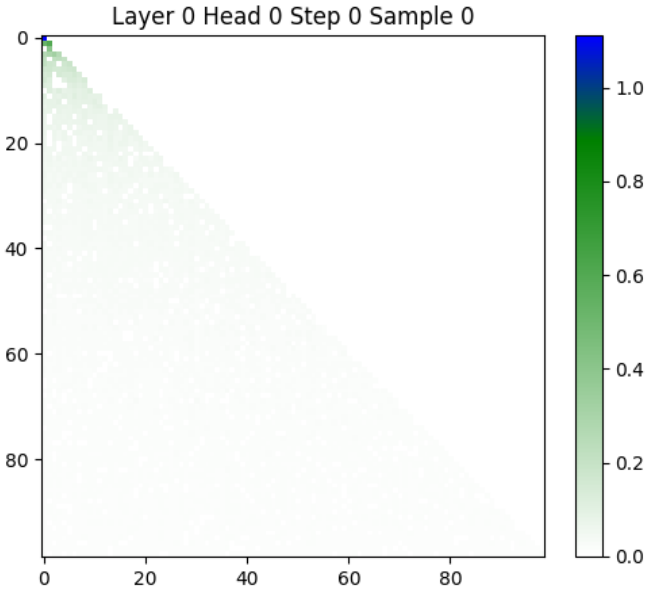
Necessary and sufficient conditions on feature learning.
b. Characterize the *path* followed by GD;
e.g., for logistic+linear, $\leq d$ "orthogonal" vecs in some geometry.

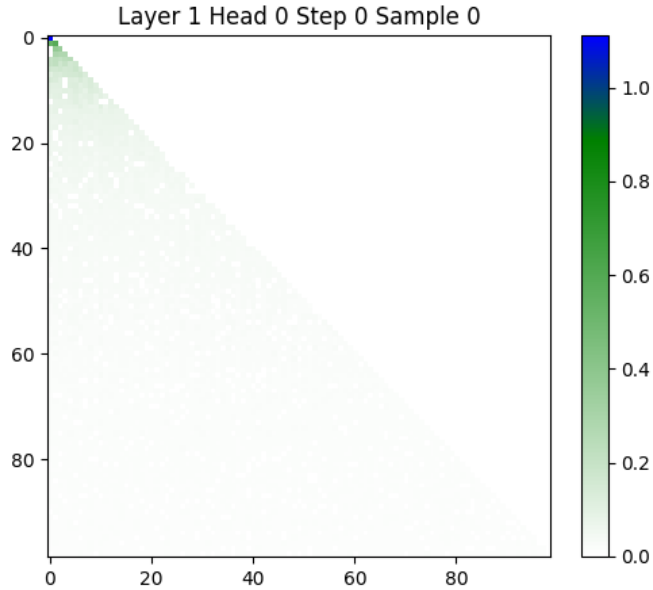
Status:

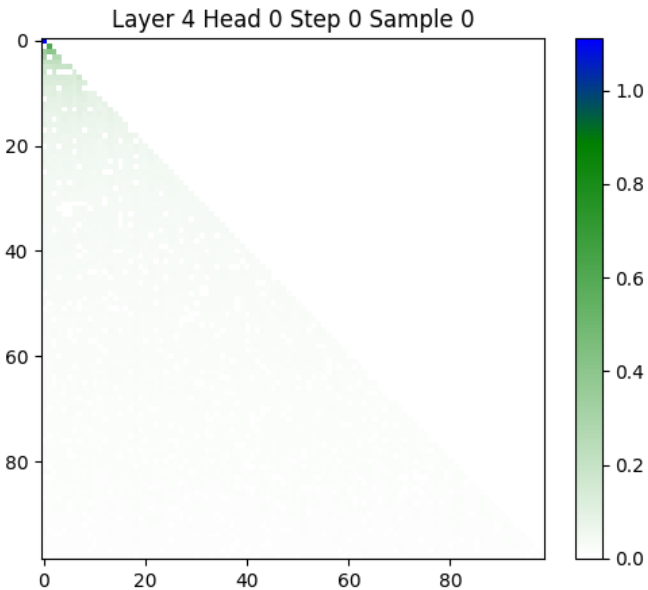
- Scattered results (Margalit Glasgow; Jingfeng Wu; . . . ?).

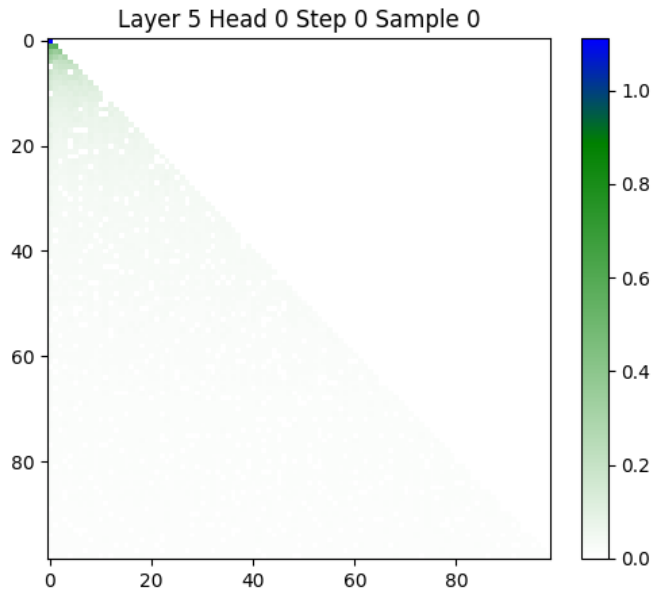
Consequences:

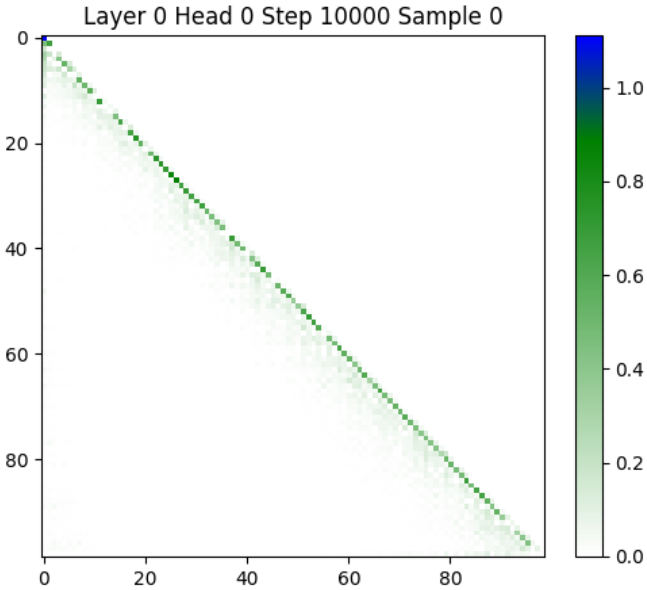
- Methodology for (≥ 2)-XOR.
- Understanding *feature diversity and coverage* in *pre-training vs post-training*.

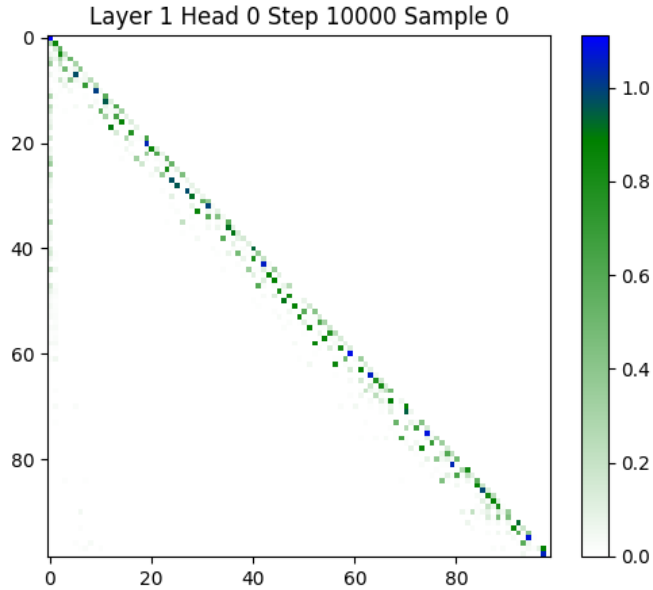

Layer 0 Head 0 Step 0 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0

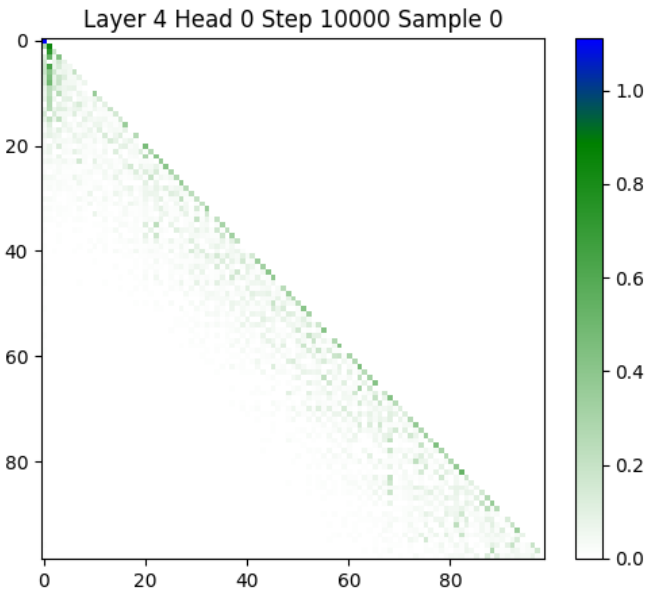

Layer 1 Head 0 Step 0 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0

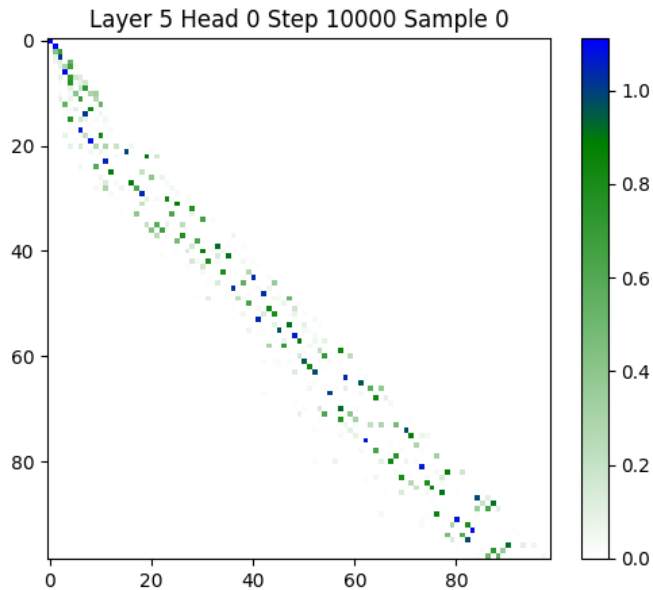

Layer 4 Head 0 Step 0 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0

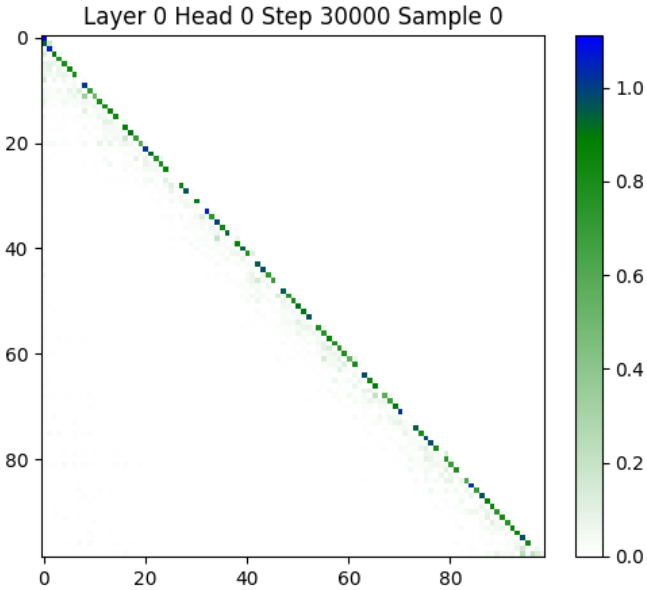

Layer 5 Head 0 Step 0 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0

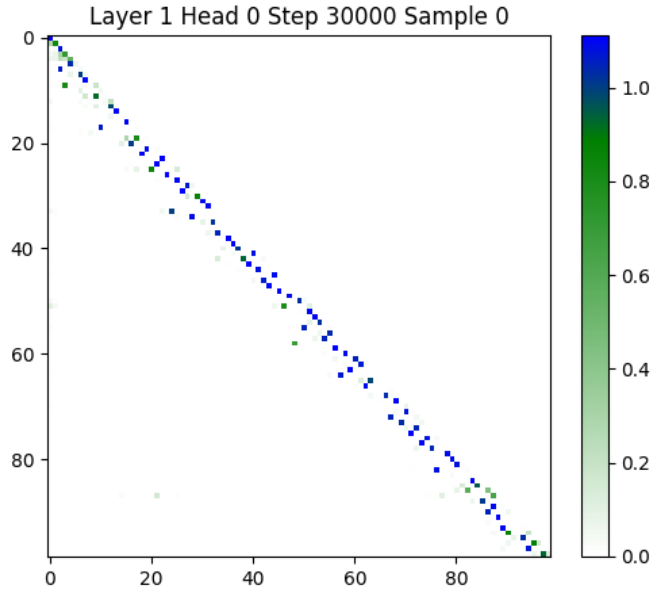

Layer 0 Head 0 Step 10000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0

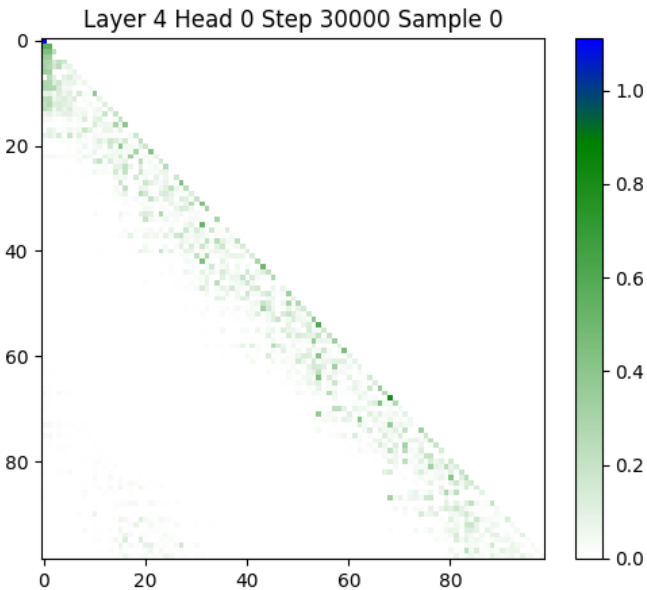

Layer 1 Head 0 Step 10000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0

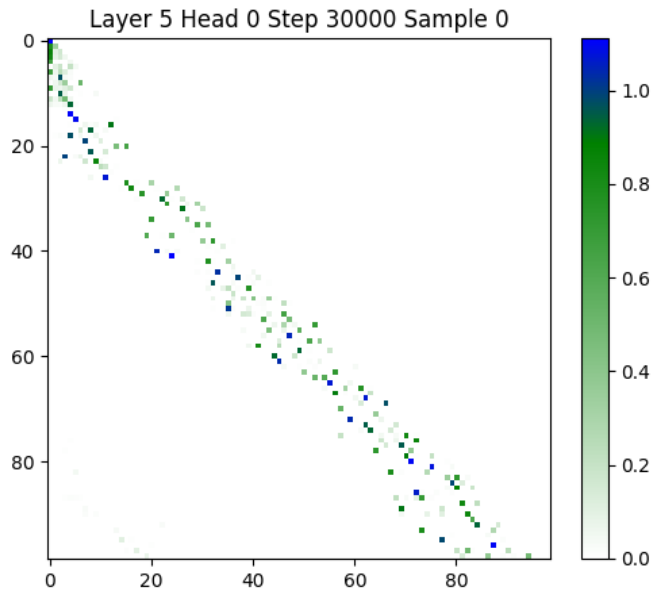

Layer 4 Head 0 Step 10000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0

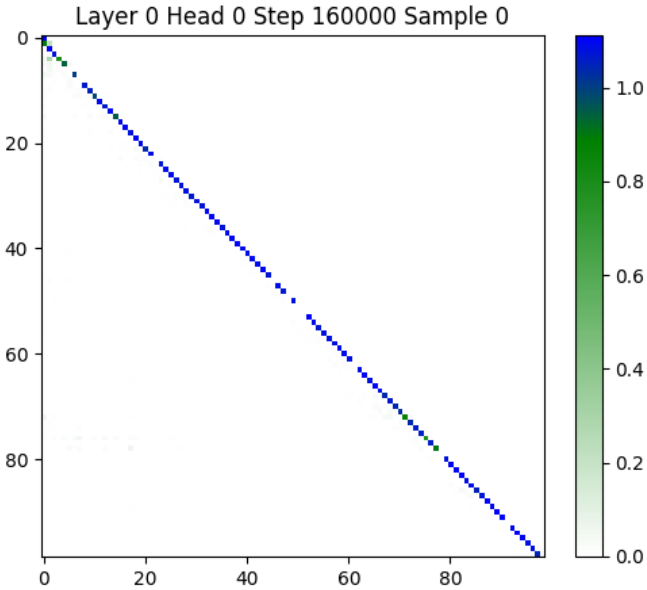

Layer 5 Head 0 Step 10000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0

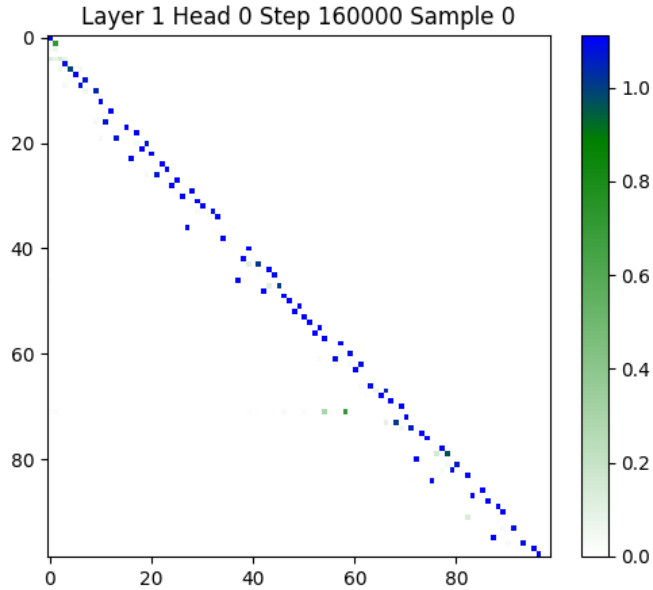

Layer 0 Head 0 Step 30000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0

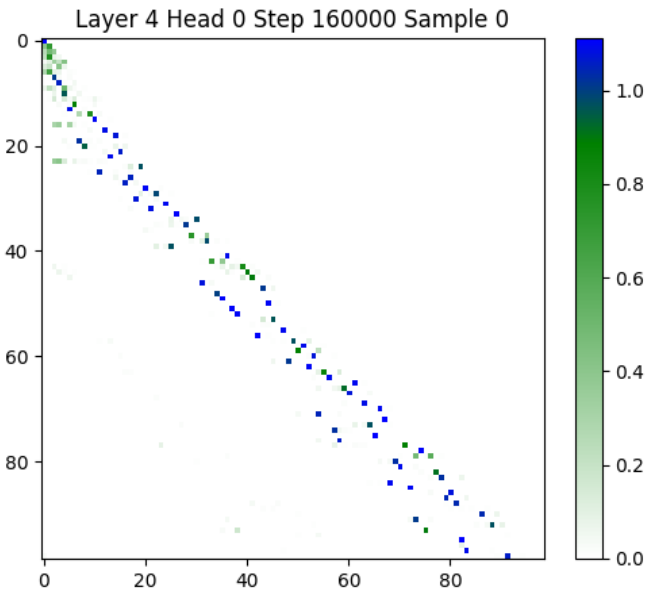

Layer 1 Head 0 Step 30000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0

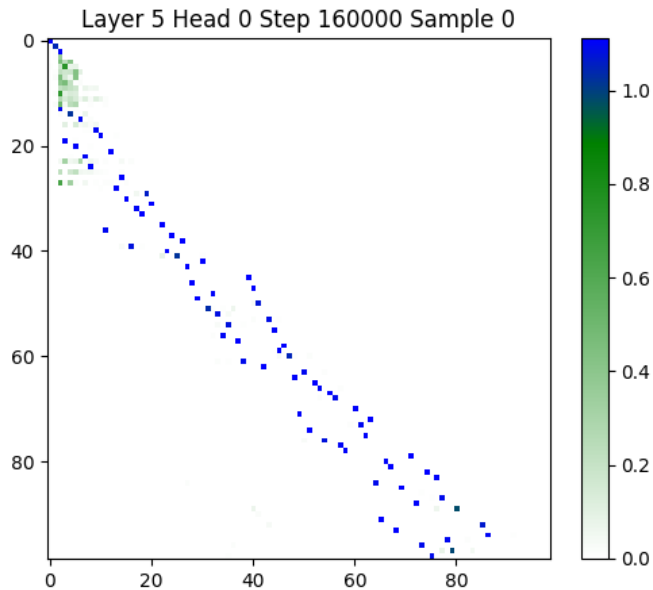

Layer 4 Head 0 Step 30000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0


Layer 5 Head 0 Step 30000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0


Layer 0 Head 0 Step 160000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0


Layer 1 Head 0 Step 160000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0


Layer 4 Head 0 Step 160000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0


Layer 5 Head 0 Step 160000 Sample 0
0
20
40
60
80
0
20
40
60
80
1.0
0.8
0.6
0.4
0.2
0.0

Open problem #1c: conditional optimization theory.

Necessary and sufficient conditions on feature learning.
c. Capture GD as an oracle:
GD in time poly$(\frac{1}{\epsilon})$ finds f with $\mathcal{L}(f) \leq \epsilon$ and $\text{Reg}(f) \leq \frac{1}{\epsilon}$.

Why we care:

- (Math) Split effort into proving the claim *or* assuming it.
- (Empirical) True inductive bias ("Reg$(\cdot)$") of GD+model.

Remarks:

- SL/RL choices: tractable/intractable $\mathcal{L}$, impractical Reg$(\cdot)$.
- A general way forward: *conditional theory*, like P $\neq$ NP.
- MLP: empirical guesses fairly widely open ($\mathcal{F}_1$ NP-hard).
- LLM: starting to understand an *algorithmic bias*.

Open ("forgotten") problem #2a: tight MLP approximation.

Empirically relevant architecture separations.
a. Theory of depth in MLPs.

Why we care:

- (Math) Pretty math is missing.
- (Empirical) Current theory provides no practical guidance.

Status:

- Some understanding for 2 vs 3 (Eldan-Shamir; Daniely; . . .).
- No understanding at higher depths
  (impractical: Telgarsky; Yarotsky; Schmidt-Hieber; . . .).

Open problem #2b: LLMs without CoT.

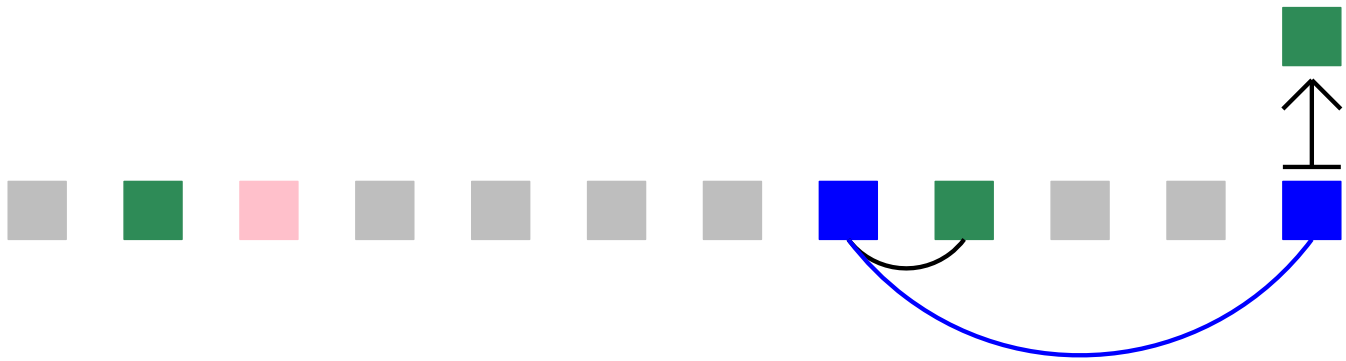
Empirically relevant architecture separations.
b. Tight computation class learned by LLMs (without CoT).

Why we care:

- (Math) a new model of computation!
- (Empirical) ideally, guidance to overcome weaknesses.

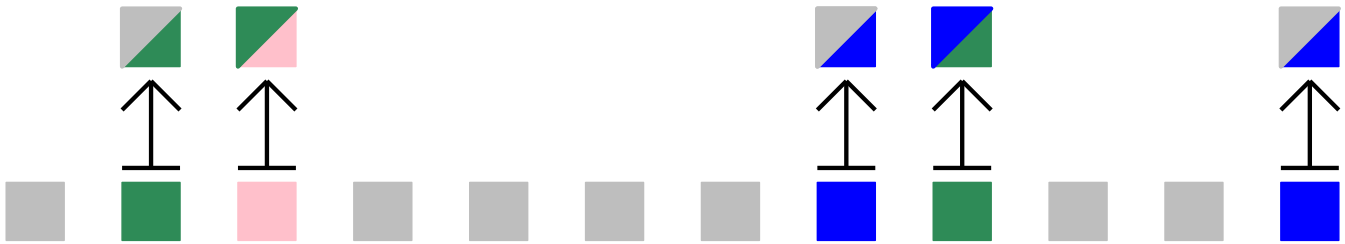
Preliminary status:

- Formalisms: C-RASP (Hahn-...),
  circuits (Merrill, Liu/Goel/Krishnamurthy/...).
- Unconditional bounds:
  (1-layer) 1-hop and Match3 (Sanford-H-T);
  ($\ln \ln n$-layer) composition task (Chen-P-W).
- *Conditional* bounds: k-hop (Sanford-H-T).

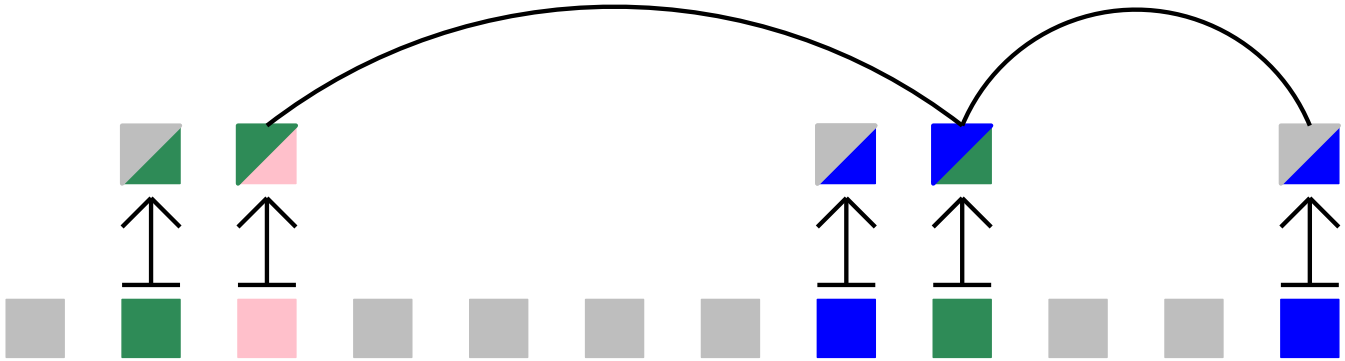


Induction heads / 1-hop (Anthropic blog, also Bietti-. . .).

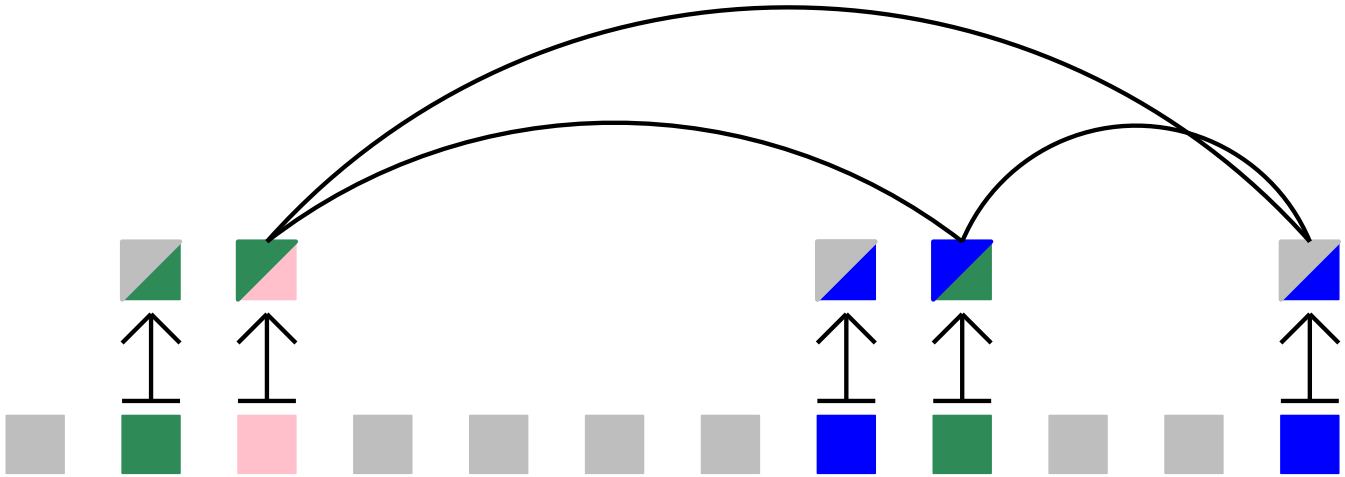
2-hop (Sanford-H-T).



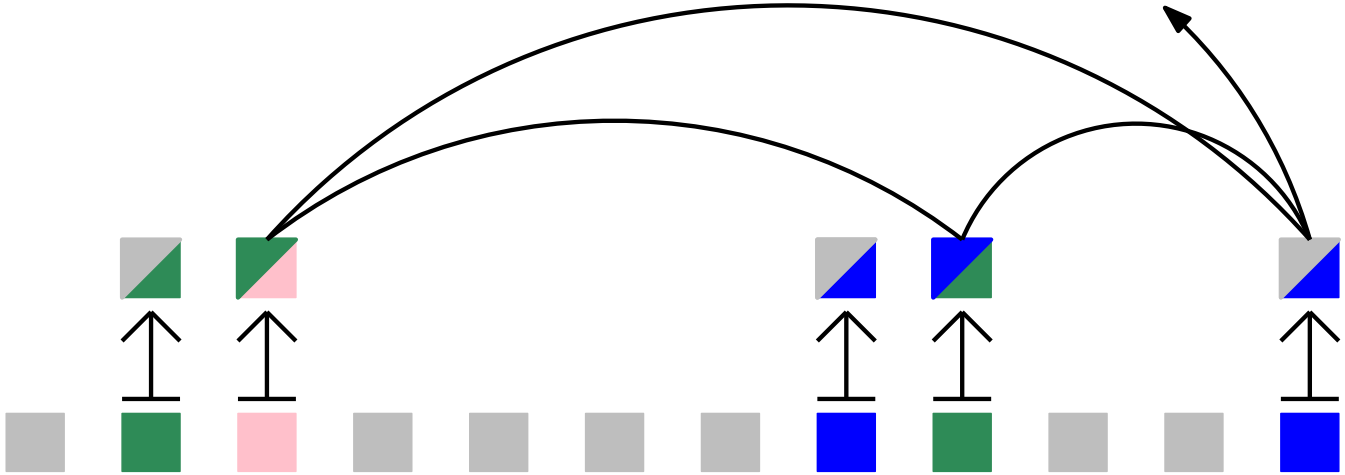
Copy forward . . .



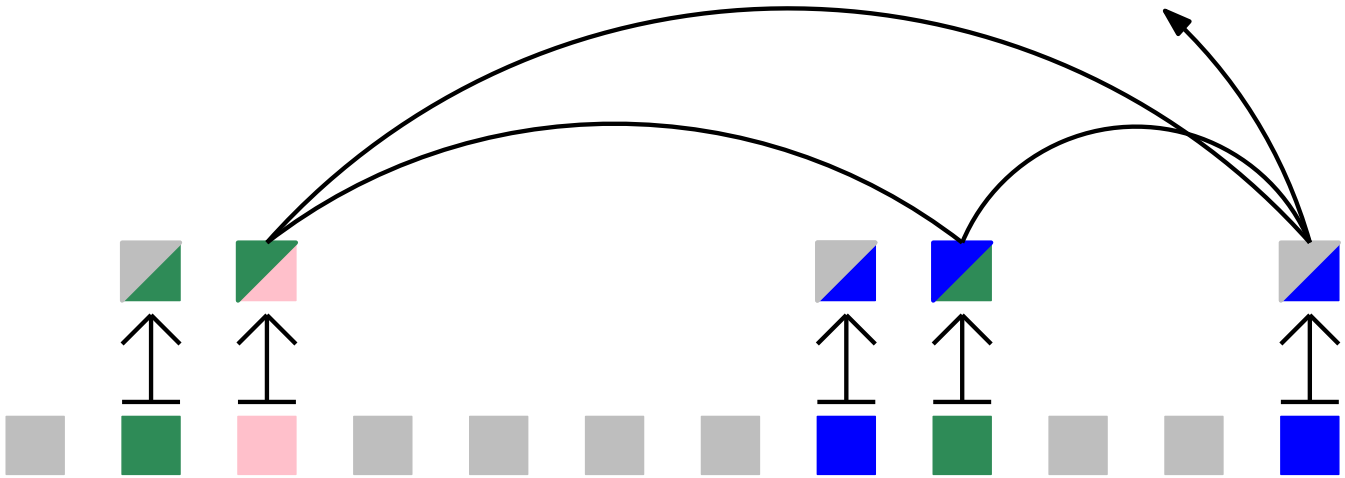
Copy forward . . . 1-hop . . .



Copy forward . . . 1-hop . . . 2-hop . . .



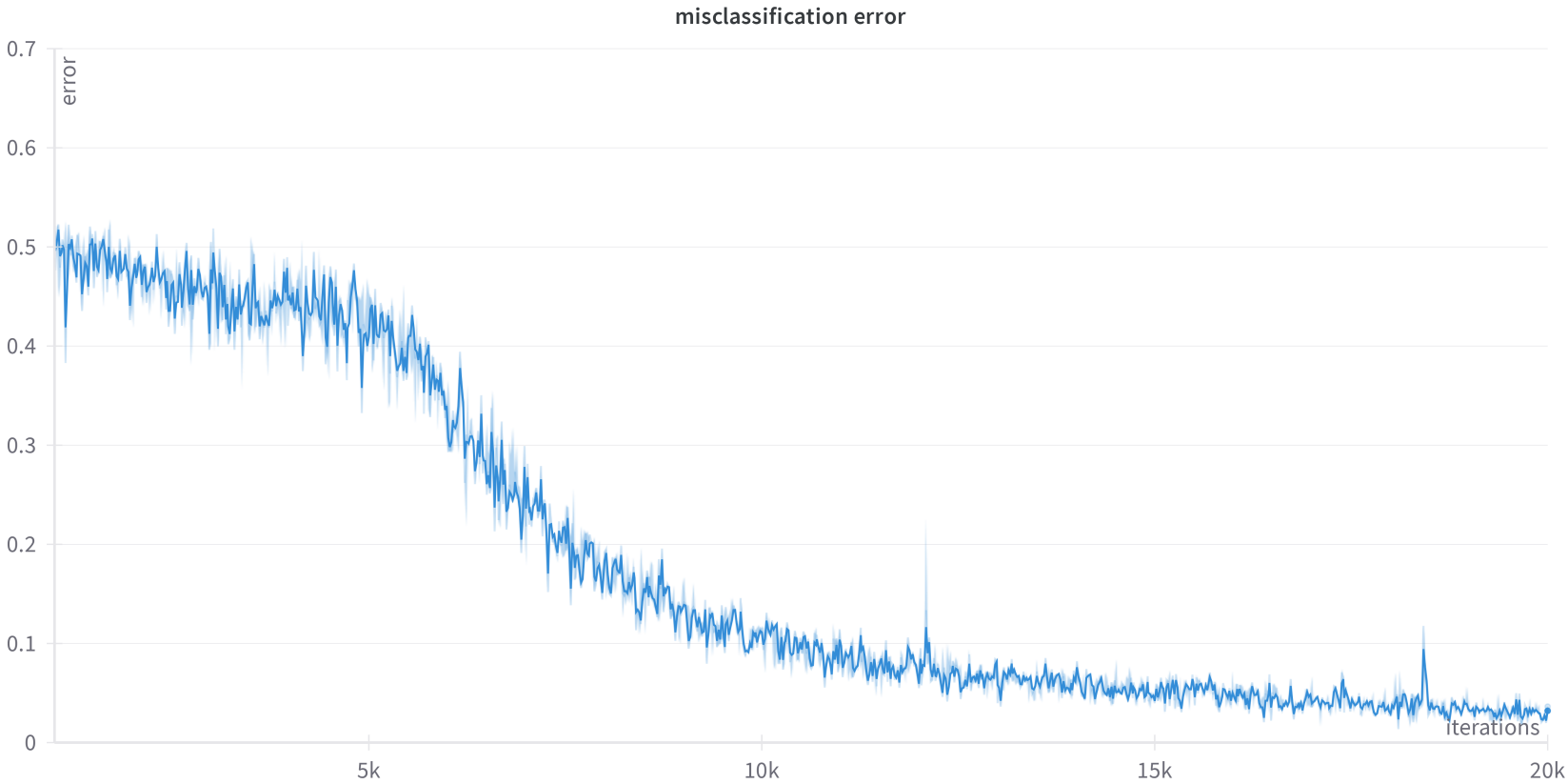
Copy forward . . . 1-hop . . . 2-hop . . . 4-hop . . .

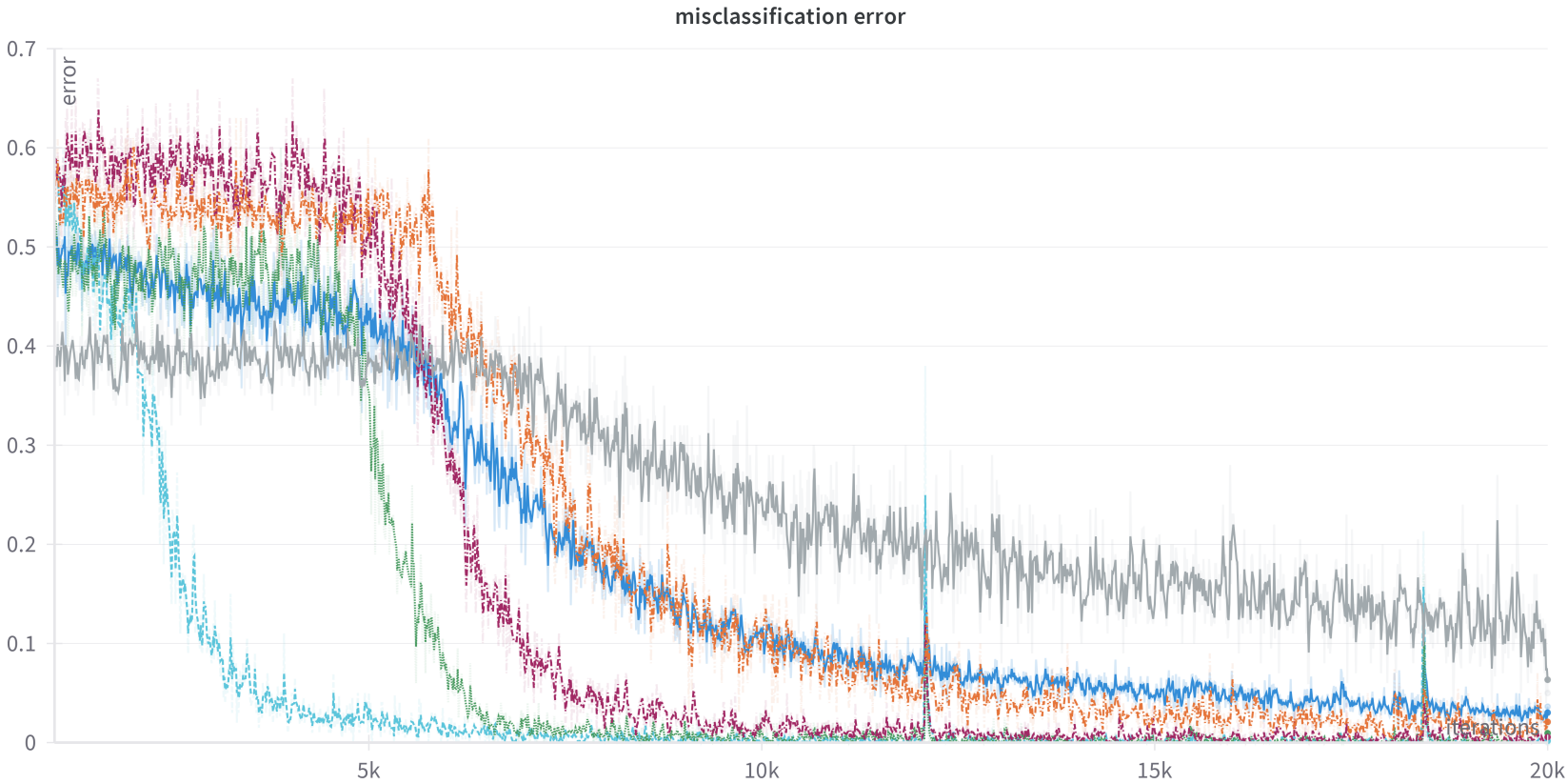


Copy forward . . . 1-hop . . . 2-hop . . . 4-hop . . .

Remarks:

- UB: $2 + \lceil \log_2 k \rceil$ layers
  (1 head, $\mathcal{O}(1)$ embedding, $\mathcal{O}(\ln n)$ precision).
- LB: *conditional* $\Omega(\ln k)$.
- Empirical learnability: next slide.
- Communication connections: 1-hop 1-layer LB, RNN LB, MPC equivalence, BAPO.


misclassification error
error
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5k
10k
15k
20k
iterations



Necessity of curriculum: Wang-N-B-D-H-L-W '25.

Open problem #2c: LLMs with CoT.

Empirically relevant architecture separations.

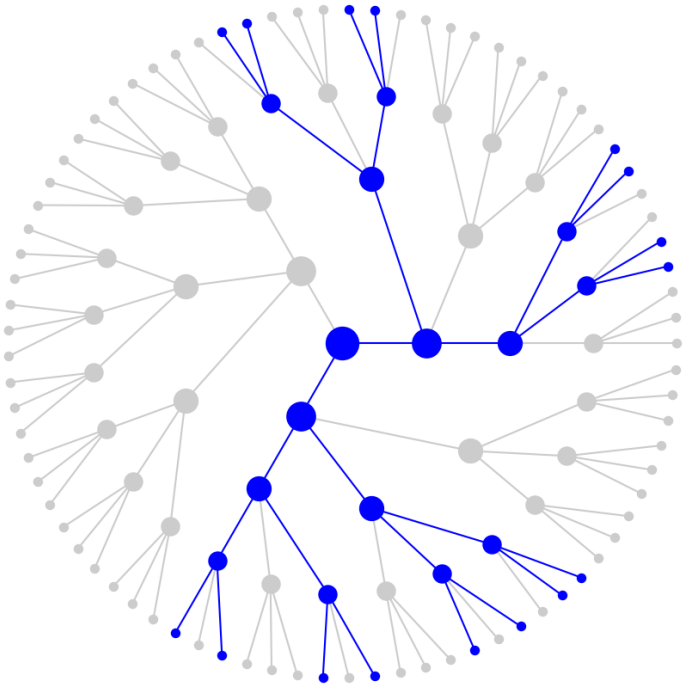
c. Pick:

- a discrete sampling problem (e.g., *next chess move*);
- abstractions of any sampling and supporting oracles (e.g., *chess move* generator, verifier, and evaluator).
- a pair of methods (e.g., LLM+CoT vs LLM+MCTS).

Establish oracle complexity separations.

Remarks:

- CoT seems a nice new sampling paradigm; not quite MCMC. (Joshi-S-V-B-G-L-M '25; Yang-S-M-L '25).
- Challenge: separate search from CoT fit to search.



$\text{depth}_{\text{low}}$	$\text{depth}_{\text{high}}$	positions	degree
1	2	29	29.00
2	3	61	7.81
4	5	175	3.64
8	11	724	2.28
16	20	44466	1.95
32	60	16917970	1.68

Challenge: LLM+CoT vs Stockfish.
How to accurately attend to millions of nodes?

Plan for today.

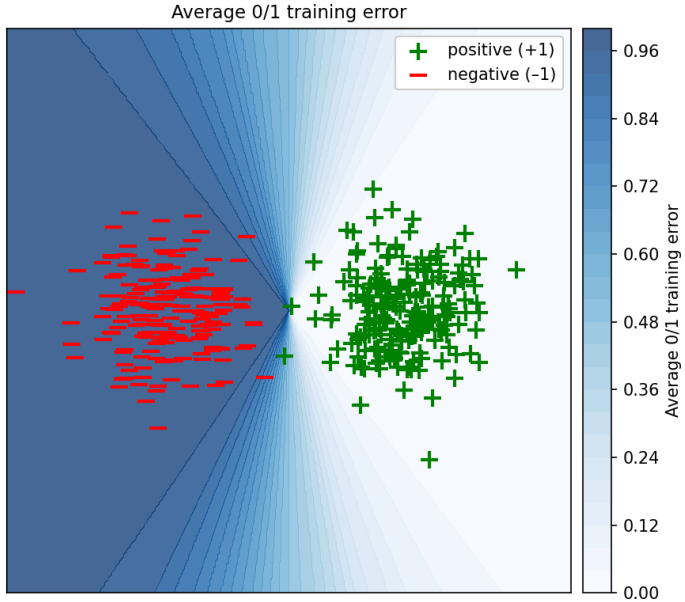
- Technical questions.
  - Feature learning with GD in MLP, LLM, *perhaps conditionally*.
  - Inductive bias of MLP, LLM, LLM+CoT.
- **Speculative questions.**
- Sociological questions.

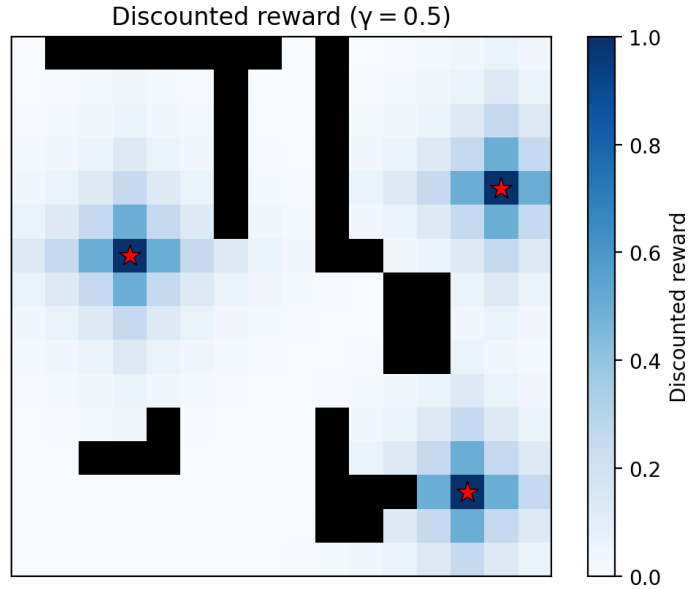
Open problem #3: universal curriculum.

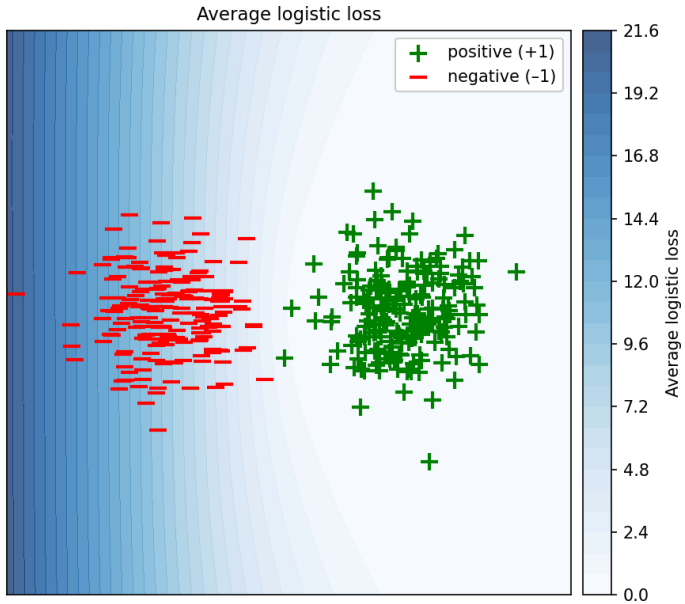
Develop a unified, effective theory of *curriculum*
aka *universal reward smoothing*.

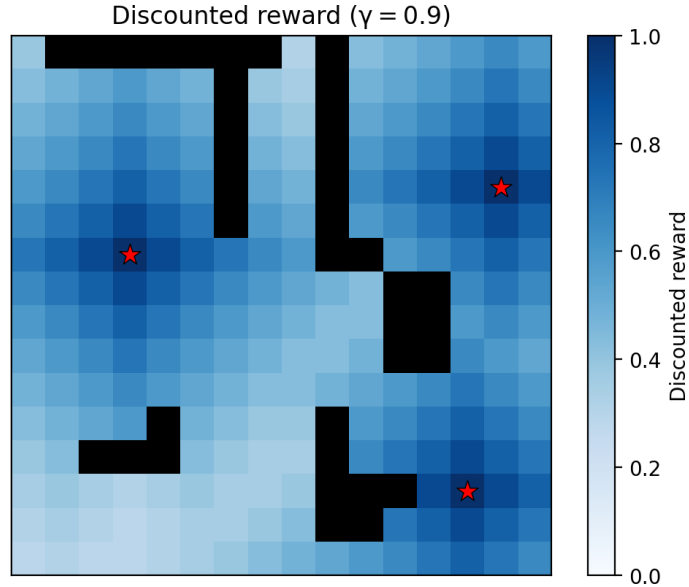
Remarks:

- Examples: surrogate losses, discounted reward, reward shaping, pre-training data proportions, CoT training data, diffusion models, ...
- *Prescriptive:* if stuck, apply curriculum somewhere.


Average 0/1 training error
positive (+1)
negative (–1)
0.96
0.84
0.72
0.60
0.48
0.36
0.24
0.12
0.00
Average 0/1 training error


Discounted reward (γ = 0.5)
1.0
0.8
0.6
0.4
0.2
0.0
Discounted reward


Average logistic loss
positive (+1)
negative (–1)
21.6
19.2
16.8
14.4
12.0
9.6
7.2
4.8
2.4
0.0
Average logistic loss


Discounted reward (γ = 0.9)
1.0
0.8
0.6
0.4
0.2
0.0
Discounted reward

Replacing the transformer.

Transformers' good news:

- Differential universal Turing machine (In n alphabet).
- Never forgets (RNN: bounded state, must commit+forget).
- Parallel computation (iterated map/reduce)
  adapted to modern hardware.

Transformers' bad news:

- "Adapted to hardware" = "hardware lottery".
- n^{th} output requires $\mathcal{O}(n)$ space, $\mathcal{O}(n)$ time.
- Hard to comprehend.

Open problem #4: replace transformer.

Extend transformer to *differentiable von Neumann architecture*:

- *instruction fetch*;
- *code is data*.

Remarks:

- "Fetch" includes formalizing RAG.
- "Code is data" includes attention outputting (Q, K, V).
- Explicit von Neumann bottlenecks.
- Core challenges:
  length generalization, effective pretraining, hardware lottery.

Open problem #5: bi-level machine learning.

Pick:

- Oracle model for training, inference, data collection, etc.
- A few *bi-level* ML models: LLM+CoT, TTT/SSL, ...

Establish oracle complexity separations.

Remarks:

- Why did bi-level structure arise now, and why LLM+CoT?
- Analogous to species/individuals and hardware/software.

Plan for today.

- Technical questions.
  - Feature learning with GD in MLP, LLM, *perhaps conditionally*.
  - Inductive bias of MLP, LLM, LLM+CoT.
- Speculative questions.
  - Curriculum learning.
  - Transformer replacement.
  - Bi-level structure.
- **Sociological questions.**

Open problem #6: the role of theory (paraphrased from Sivaraman Balakrishnan).

Open problem #6: the role of theory (paraphrased from Sivaraman Balakrishnan).

Clarify and communicate the roles of theory:

1. Analysis of existing algorithms, mental models, empirical rules-of-thumb. (Accept being left behind.)
2. Design of new algorithms. (E.g., for test-time inference, safety, interpretability, . . .)
3. Rigor and pretty mathematics. (Enjoy being behind.)

Suggestions:

- Senior (those with hiring power): make hiring/promotion requirements explicit. Junior: hedge 😵, absorb new tools.
- Conferences shield less popular topics and slower timelines.
- Adopt GPUs ubiquitously (e.g., $1000 experiment or proof?).

Open problem #7: the role of open science.

Why didn't academia discover "small-scale o1";
e.g., LLM+CoT not LLM+MCTS for chess?

Remarks:

- o1 weds all AI history: search, RL, opt, ...
- Concern: academia is slow, risk-averse, closed, complacent (especially with teaching!).
  E.g., where to publish speculative theories of generalization?
- Publication = 1800s pen-and-paper *vs* global public git.
  ("Alpaca lesson".)

Plan for today.

- Technical questions.
  - Feature learning with GD in MLP, LLM, *perhaps conditionally*.
  - Inductive bias of MLP, LLM, LLM+CoT.
- Speculative questions.
  - Curriculum learning.
  - Transformer replacement.
  - Bi-level structure.
- Sociological questions.
  - How to re-align and re-incentivize theory?
  - How to make open science more open?

Thank you!

Hello, Friend.