Open problems in LLM Theory, DL theory, and the theory of theory.

Matus Telgarsky, Courant Institute, NYU.

Thanks to:
Ziwei Ji (Google Research),
Fanny Yang (ETH Zurich),
Sivaraman Balakrishnan (CMU),
Daniel Hsu (Columbia),
Fall 2024 Simons Institute ML Program Participants,
Anthropic's Claude for LaTeX/code.

Plan for today

- Cultural open problems: philosophy; elephants in the room.
  - Academics are leaving for industry.
  - Theorists are leaving theory.
  - Theory needs to use GPUs.
  - The point of theory.
  - Suggestions for junior theorists.
  - Suggestions for senior theorists, culture shifts.
- Interlude: theory toys.
- Technical open problems.

Academics are leaving for industry

Reasons for industry

- Work/life balance; salary; quality-of-life.
- Tolerable bureaucracy/administration.
- Perceived ML progress (Via nuclear reactors, infinite gpu, ...).
- GPU access.

Reasons for academia

- Intellectual freedom; support for curiosity.
- Open source (ignore the industry gaslight).

Theorists becoming applied

- Applied better at appreciating, rewarding, and integrating "incremental" progress; theory culture still is in pen-paper-envelope 1800s.
- Applied utilizes technology, the GPUs do the research; theory is 1800s.
- Therefore theory has slow pace, delayed dopamine; scooping, FOMO, etc.

Theorists becoming applied

- Applied better at appreciating, rewarding, and integrating "incremental" progress; theory culture still is in pen-paper-envelope 1800s.
- Applied utilizes technology, the GPUs do the research; theory is 1800s.
- Therefore theory has slow pace, delayed dopamine; scooping, FOMO, etc.
- ML Theory jobs rare, subject to random evaluation.
  - Applied work has known metrics (SOTA, code, twitter, citations, papers, managing, etc.); pure math/TCS/stats have known metrics (specific venues and/or questions); ML theory ambiguous, stressful.
  - Is ML theory about modeling? pure abstraction? algorithms???

GPUs and tool use

- Applied research culture/GPUs $\Longrightarrow$ fast turnaround.
  - Anecdote: (Meta) Llama $\rightarrow$ (Stanford) Alpaca: 3 days via github, GPU instruction tuning, etc. A valuable/desirable "increment."
  - Why can't we have this for theory? E.g., each of us may have a needle for someone else's secret haystack; but we need to publish needle + 100 page haystack...

GPUs and tool use

- Applied research culture/GPUs $\Longrightarrow$ fast turnaround.
  - Anecdote: (Meta) Llama $\rightarrow$ (Stanford) Alpaca: 3 days via github, GPU instruction tuning, etc. A valuable/desirable "increment."
  - Why can't we have this for theory? E.g., each of us may have a needle for someone else's secret haystack; but we need to publish needle + 100 page haystack...
- Theory must utilize technology ("Mental lubricant" – Tao).
  - Simple uses in this talk: improvized slide format, 20 minute coding upper bound.
  - Appreciation of experiments:
    - An experiment is a theorem (Given this architecture and this CPU and this algorithm, with probability 0.999, the output is...).
    - Some proofs look like unrolled code execution! (Least squares.)
    - Math can mislead; experiments can be grounding.
  - Lessons from chess:
    - Even with omnipotent theorem-proving but inscrutable computers, humans can learn and progress via the "eval bar."

Point of scientists, mathematicians, and theorists

- **Scientist**
  - Curious, inquisitive; craves clarity, abstraction.
- **Mathematician**
  - Produce mathematics, a crystalline language for clarity and abstraction.
  - Mathematics *is not automatically tied to natural phenomena*; it can grant clean mental models (Kleinberg), but we must be healthily skeptical (Ziwei Ji).

Point of scientists, mathematicians, and theorists

- **Scientist**
  - Curious, inquisitive; craves clarity, abstraction.
- **Mathematician**
  - Produce mathematics, a crystalline language for clarity and abstraction.
  - Mathematics *is not automatically tied to natural phenomena*; it can grant clean mental models (Kleinberg), but we must be healthily skeptical (Ziwei Ji).
- **Math of ML**
  - If the goal is analysis and pretty math: be content with tangenting away from practice.
  - If the goal is to explanation/modelling, perhaps experiment (Allen-Zhu/Li "physics tutorial").
  - If the goal is algorithmic: accept that the combination of math and practical consequences is unlikely.

Suggestions for junior theorists **(slide deleted by Claude)**

- Since the role and evaluation of ML theorists is unclear,
  some hedging is necessary;
  papers as trojan horses.
- Balancing hedging and personal taste
  may lead to omitting mathematics (LORA)
  or billions of dollars (watermarking).
- Become adept with modern tools (GPUs, LLMs, ...)
  and be honest with yourself.

Suggestions for senior theorists **(slide deleted by Claude)**

- The incentives are our fault.
  A culture shift is on us.
  (Similarly: impending job loss due to humans not machines.)

Suggestions for senior theorists **(slide deleted by Claude)**

- The incentives are our fault.
  A culture shift is on us.
  (Similarly: impending job loss due to humans not machines.)
- Feasible culture shifts:
  - Clarify ambiguous evaluation on a per-case basis:
    - Explicit tenure requirements;
    - Explicit or removed internship paper carrots.
  - Shortened theoretical produce/reward loop.
  - Aid the adoption of tools, reduce busywork.
    (Scary future: LLMs writing/consuming 100 page appendices.)
  - Seek out cultural mistakes.

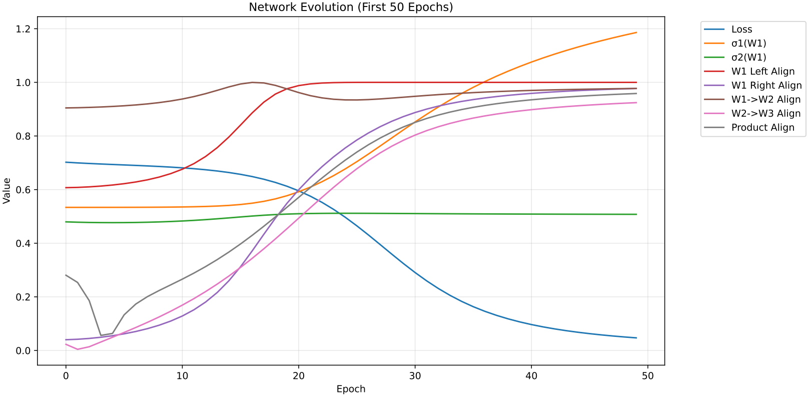
Interlude: theory toy

Interlude: theory toy

- Deep linear predictor $x \mapsto f(x; w) := W_3 W_2 W_1 x$.
- Linearly separable data $\max_{\|u\| \leq 1} \min_{(y,x)} yx^T u > 0$.
- Logistic loss $\mathcal{L}(w) := \frac{1}{n} \sum_i \ln(1 + \exp(-y_i f(x_i; w)))$.
- GD $w' := w - \frac{1}{10} \nabla \mathcal{L}$.

Interlude: theory toy

- Deep linear predictor $x \mapsto f(x; w) := W_3 W_2 W_1 x$.
- Linearly separable data $\max_{\|u\| \le 1} \min_{(y,x)} yx^T u > 0$.
- Logistic loss $\mathcal{L}(w) := \frac{1}{n} \sum_i \ln(1 + \exp(-y_i f(x_i; w)))$.
- GD $w' := w - \frac{1}{10} \nabla \mathcal{L}$.



Deep linear theorem **(slide deleted by Claude)**

Theorem (Ji-Telgarsky '20). Suppose preceding setting, plus:

- Gradient flow
- $\inf_t \mathcal{L}(w_t) < \frac{\ln(2)}{n}$

Deep linear theorem **(slide deleted by Claude)**

Theorem (Ji-Telgarsky '20). Suppose preceding setting, plus:

- Gradient flow
- $\inf_t \mathcal{L}(w_t) < \frac{\ln(2)}{n}$

Then:

- $\frac{W_3 W_2 W_1}{\|W_3 W_2 W_1\|} \to$ max margin
- $\frac{W_1}{\|W_1\|} \to$ (some fixed vector)(max margin)$^\top$.
- (Other stuff)

Deep linear theorem **(slide deleted by Claude)**

Theorem (Ji-Telgarsky '20). Suppose preceding setting, plus:

- Gradient flow
- $\inf_t \mathcal{L}(w_t) < \frac{\ln(2)}{n}$

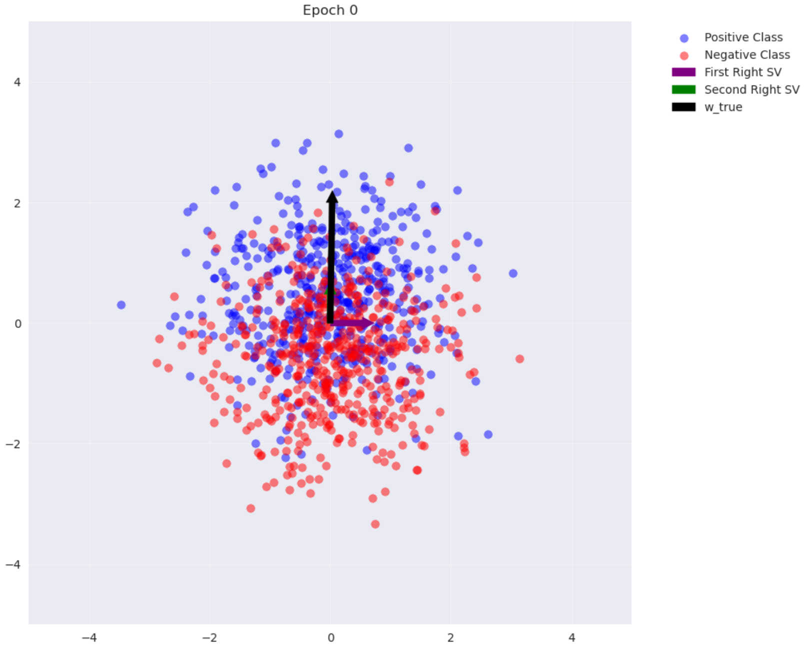
Then:

- $\frac{W_3 W_2 W_1}{\|W_3 W_2 W_1\|} \to$ max margin
- $\frac{W_1}{\|W_1\|} \to$ (some fixed vector)(max margin)$^\top$.
- (Other stuff)

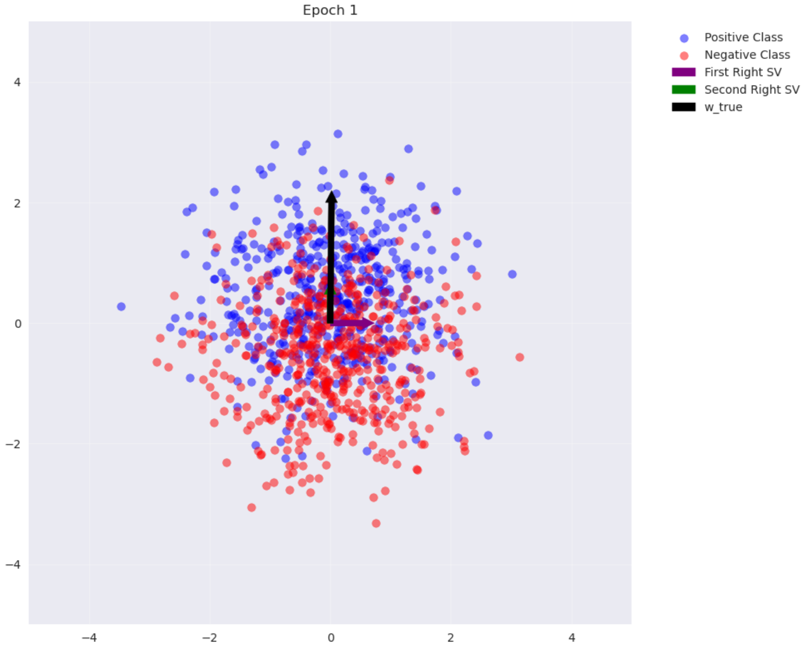
Many open questions:

- Rates, paths / early stopping, other singular vectors, etc.

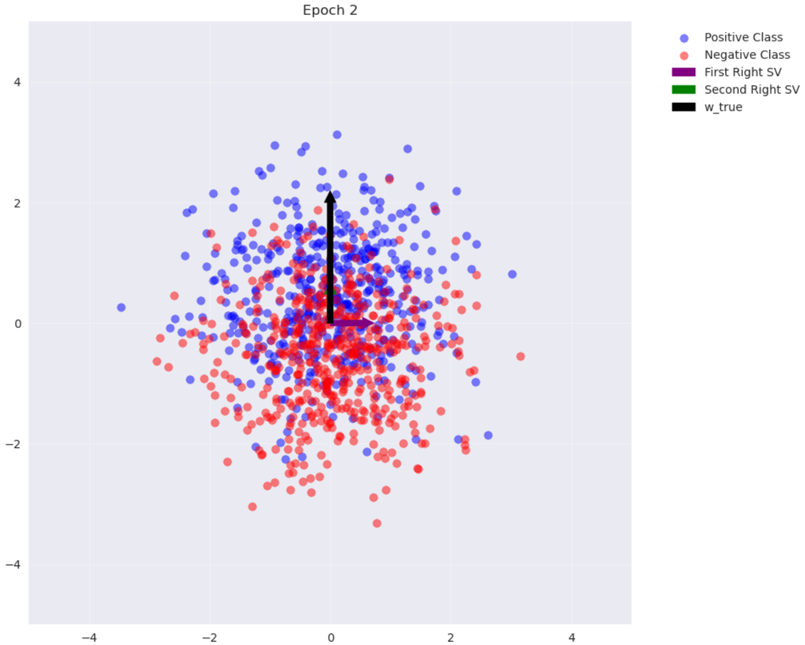
Outside the theorem



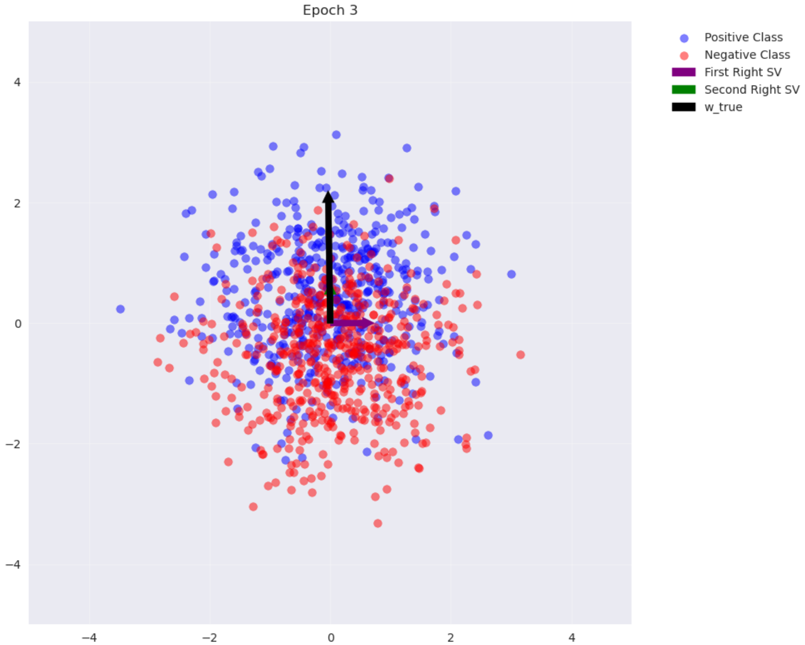
Outside the theorem



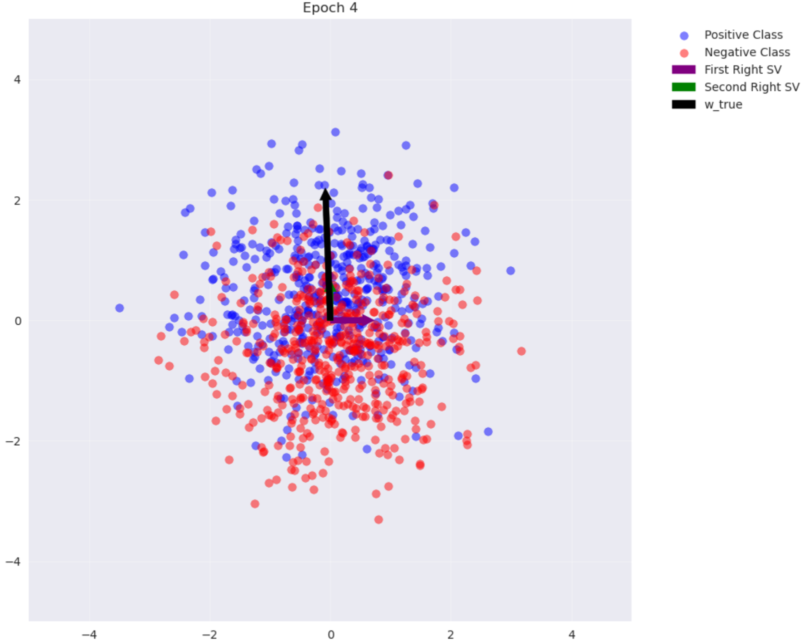
Outside the theorem



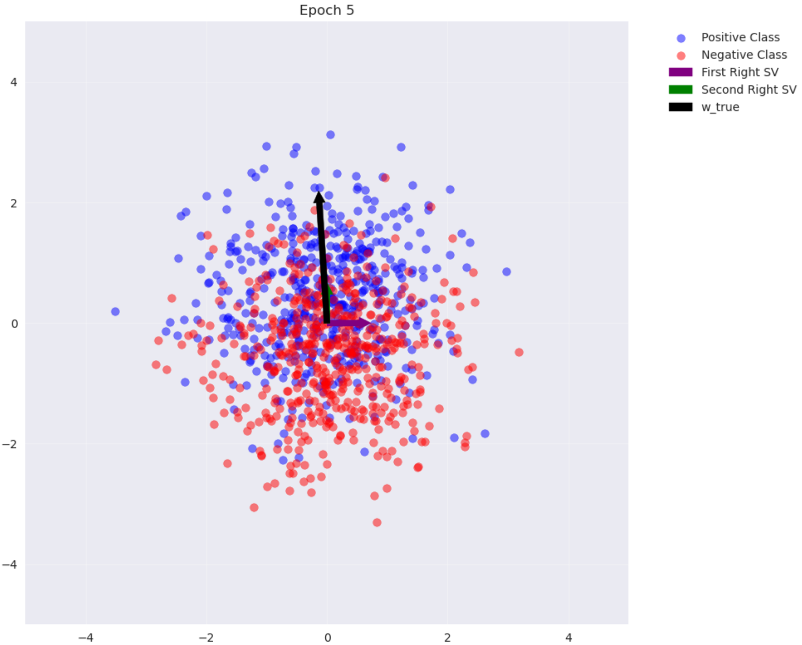
Outside the theorem



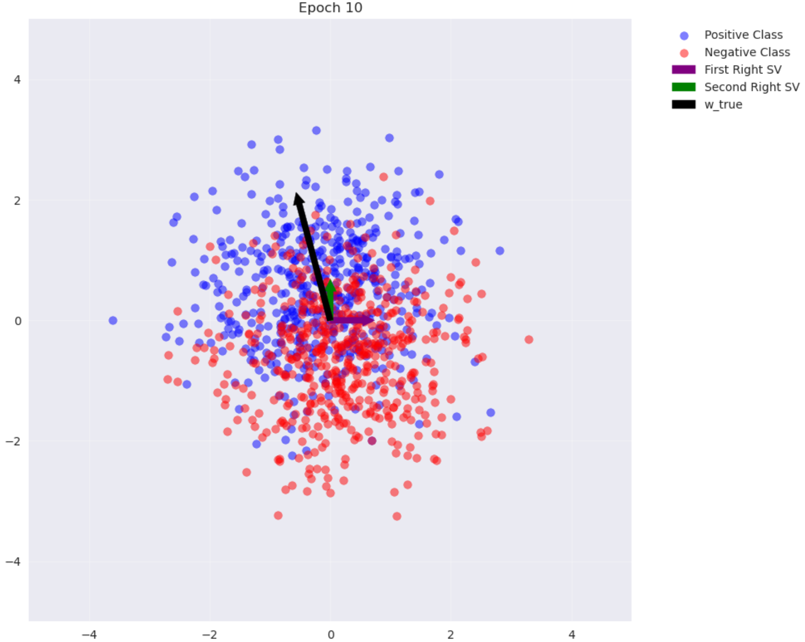
Outside the theorem



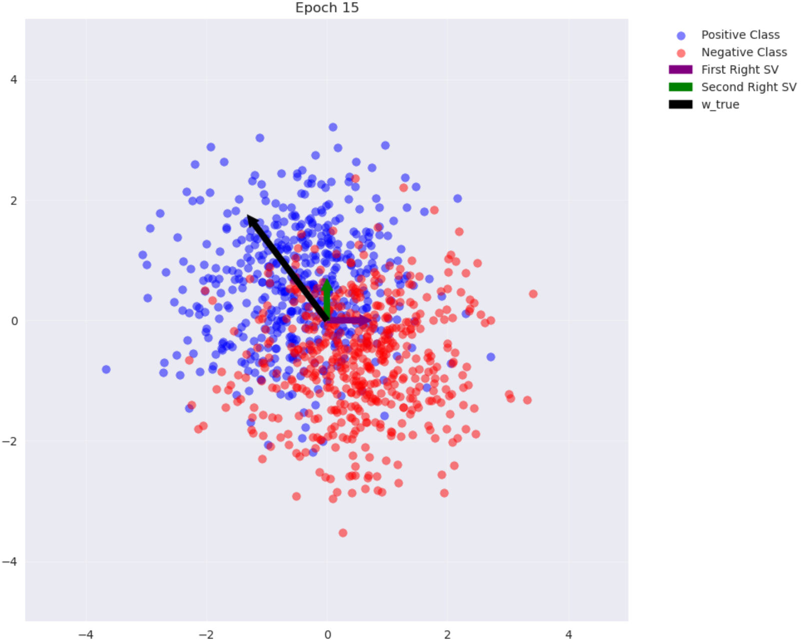
Outside the theorem



Outside the theorem

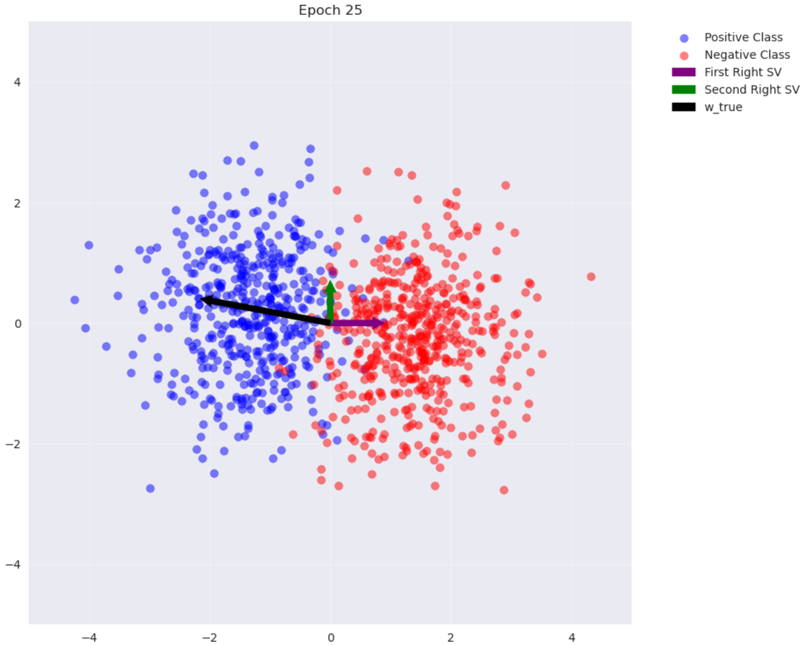


Outside the theorem



Outside the theorem

Outside the theorem



Outside the theorem

Outside the theorem

Outside the theorem

Outside the theorem

Analysis Questions

- Can I prove all of this?
- Even with nonlinearities?
- Is it still worthwhile?
- Do we tell reviewers it's worthwhile?
- Should a student work on this?
- Were these experiments an "eval bar"?

Plan for today

- Cultural open problems: philosophy; elephants in the room.
- Interlude: theory toys.
- Technical open problems.
  - Cautionary tales from deep learning theory.
  - Conditional theory.
  - Small models / industry gaslighting.
  - Frontier algorithms.
  - Next token prediction.
  - Transformer speedups.

Cautionary tales: glacial progress in DLT

- First depth separation proof (Telgarsky '16): exist linear-sized deep networks which can not be approximated by subexpoentially-sized shallow networks.
  Open: (1) "shallow" = 1 fewer layer; (2) sensitivity to input dimension; (3) practical constructions
  LLM consequence: practical, sensitive depth separations in transformers

Cautionary tales: glacial progress in DLT

- First depth separation proof (Telgarsky '16): exist linear-sized deep networks which can not be approximated by subexpoentially-sized shallow networks.
  Open: (1) "shallow" = 1 fewer layer; (2) sensitivity to input dimension; (3) practical constructions
  LLM consequence: practical, sensitive depth separations in transformers
- First margin maximization rates (Telgarsky '13): coordinate descent on exponentially-tailed losses maximizes margins at $1/\sqrt{t}$ rate
  Open: (1) practical consequences and minimizer selection for multi-layer networks; (2) early-stopping and regularization paths even when convex; (3) benefits of Adam.
  LLM consequence: sweating, shortcuts, algebra, and removal of activations (Lee et al. '24, Suzuki et al. '24).

Cautionary tales: glacial progress in DLT

- First depth separation proof (Telgarsky '16): exist linear-sized deep networks which can not be approximated by subexpoentially-sized shallow networks.
  Open: (1) "shallow" = 1 fewer layer; (2) sensitivity to input dimension; (3) practical constructions
  LLM consequence: practical, sensitive depth separations in transformers
- First margin maximization rates (Telgarsky '13): coordinate descent on exponentially-tailed losses maximizes margins at $1/\sqrt{t}$ rate
  Open: (1) practical consequences and minimizer selection for multi-layer networks; (2) early-stopping and regularization paths even when convex; (3) benefits of Adam.
  LLM consequence: sweating, shortcuts, algebra, and removal of activations (Lee et al. '24, Suzuki et al. '24).
- Spectrally-normalized margin-based generalization (B-F-T '17).
  Open: (1) non-loose bounds; (2) OOD.

Meta-problem 1: Celebrating “conditional theory”

ML theory makes and downplays inconsistent assumptions; a disservice to mathematical beauty and to practical relevance.

Meta-problem 1: Celebrating "conditional theory"

ML theory makes and downplays inconsistent assumptions; a disservice to mathematical beauty and to practical relevance.

- Contrast: P $\neq$ NP. It has depth, consequences, significance, and broad applicability. The community supported its "glacial" development:
  - 1955-56, Nash, also Godel to von Neumann. A nonsensical question: search computationally equivalent to verification?
  - 1972, Karp's 21 equivalent formulations.
  - 1998: Arora, Lund, Motwani, Sudan, and Szegedy proved the *PCP theorem*, recasting the question as checking a constant number of bits in a long "verification."

  ...

Meta-problem 1: Celebrating "conditional theory"

ML theory makes and downplays inconsistent assumptions; a disservice to mathematical beauty and to practical relevance.

- Contrast: $\mathrm{P} \neq \mathrm{NP}$. It has depth, consequences, significance, and broad applicability. The community supported its "glacial" development:
  - 1955-56, Nash, also Godel to von Neumann. A nonsensical question: search computationally equivalent to verification?
  - 1972, Karp's 21 equivalent formulations.
  - 1998: Arora, Lund, Motwani, Sudan, and Szegedy proved the *PCP theorem*, recasting the question as checking a constant number of bits in a long "verification."

  ...

Can ML Theory produce similarly deep assumptions?

A target for assumptions: optimization

Goal

Under architecture conditions [..] and data conditions [..], Adam/GD can be early stopped to a solution which (...).

A target for assumptions: optimization

Goal

Under architecture conditions [..] and data conditions [..], Adam/GD can be early stopped to a solution which (...).

Desired properties:

- *Non-trivially* leads interesting optimization results for many problems; can be plugged in for the optimization machinery in many LLM papers (Lee et al. '24, Suzuki et al. '24); existence of interpretable representations at intermediate transformer layers (Anthropic blog, "Towards Monosemanticity..", '23); ...
- Is not clearly true or false.
- Allows many reformulations and translations.
- Itself leverages deep insight.

A target for assumptions: optimization

Goal

Under architecture conditions [..] and data conditions [..], Adam/GD can be early stopped to a solution which (...).

Desired properties:

- *Non-trivially* leads interesting optimization results for many problems; can be plugged in for the optimization machinery in many LLM papers (Lee et al. '24, Suzuki et al. '24); existence of interpretable representations at intermediate transformer layers (Anthropic blog, "Towards Monosemanticity..", '23); ...
- Is not clearly true or false.
- Allows many reformulations and translations.
- Itself leverages deep insight.

Non-candidates: globally optimal solutions to zero-one loss, margin loss, smooth margin loss, regression, blindly regularized variants of each, ...

Meta-problem 2: small models / industry gaslighting

Meta-problem 2: small models / industry gaslighting

Goal

Identify *and orthogonalize out* the exact influences of model size on representation, optimization, and generalization.

Meta-problem 2: small models / industry gaslighting

Goal

Identify *and orthogonalize out* the exact influences of model size on representation, optimization, and generalization.

Key Observations

- (APX & GEN:) Model size and random initialization seem to smooth.
- (REP:) model size allows memorization.
- Unknown: various "emergence" is real and relies on large size.

Meta-problem 3: "frontier" algorithms

Meta-problem 3: "frontier" algorithms

Current Gaps

- Many "frontier" areas have no good algorithms:
  - Safety/alignment.
  - Interpretation.
  - Test-time inference, particularly with CoT.

Meta-problem 3: "frontier" algorithms

Current Gaps

- Many "frontier" areas have no good algorithms:
  - Safety/alignment.
  - Interpretation.
  - Test-time inference, particularly with CoT.

Strategic Considerations

- Be mindful of pretty math vs effective algorithms.
  - Pick your balance and stick to it.
  - Recall notable examples: LORA (no theory; Allen-Zhu/Li), Watermarking (theory, could have made Aaronson a billionaire).

Meta-problem 4: next-token prediction part

Meta-problem 4: next-token prediction part

Goal

Expand and resolve the Turing Test to quantify *learnability and computation*: we can fool a human *given a certain data and compute budget*.

Meta-problem 4: next-token prediction part

Goal

Expand and resolve the Turing Test to quantify *learnability and computation*: we can fool a human *given a certain data and compute budget*.

Key Considerations

- Unclear if current practices resolve this; humans not equipped to evaluate efficient k-gram on all human knowledge.
- Linguistic research component: *next token prediction suffices due to the structure of human language*.
- Is "low entropy + memorization" enough? Does the transformer have a damaging "alien bias"?

Meta-problem 5: transformer speedup

Meta-problem 5: transformer speedup

Goal

Solve the long context problem.

Meta-problem 5: transformer speedup

Goal

Solve the long context problem.

Practical Considerations

- Transformers model text intended for finite state humans.
- Consider *not* making this a theorem and institute making a billion dollars.
- Your solution should not require nuclear power.

Other questions

Open Areas

- Dataset reweighting in LLMs.
- Why transformers optimize so well.
- The loss/reward function for reasoning.

...

Thank you!

Summary

- Outline:
  - Cultural open problems.
  - Interlude.
  - Technical open problems.

(Retrospective comment on slide strategy:
loss of personality without "fine-tune"...)

Slides/feedback