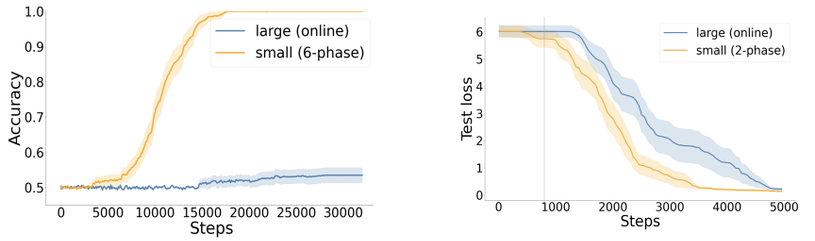
Less Data, Faster Training

Repeating smaller datasets speeds up learning via sampling biases

Jingwen Liu, Ezra Edelman, Surbhi Goel, Bingbin Liu

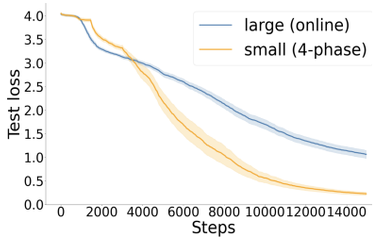
Columbia University · Penn (University of Pennsylvania) · Kempner Institute

Small-vs-large gap: less data → better performance, under the same compute.

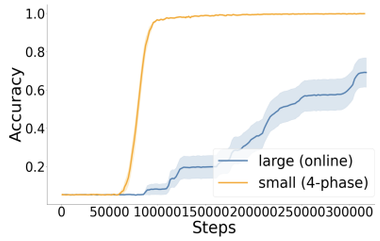


(a) sparse parity

(b) single-index model



(c) in-context linear regression



(d) modular addition

TL;DR: The small-vs-large gap can be explained by the **relative norm adjustments via sampling biases**.

- Intuition: sampling biases grow … formalized by 2-layer MLP on sparse parity.
- Empirically validated with 1) random label training; 2) layerwise treatments.

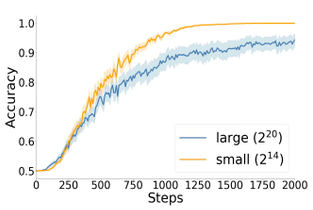
∴ Repetition as favorable inductive biases for optimization, e.g. LLM post-training on reasoning tasks.

Case study: (d, k)-sparse parity.

$$x = 1\ {-1}\ \underbrace{{-1}\ 1\ {-1}}_{S}\ 1\ 1\ 1\ {-1}\ 1 \in \mathbb{R}^d \rightarrow y = \prod_{i \in S} x_i = 1$$

Prior theory is insufficient.

- CSQ-SQ separation … coincide.
- Gradient variance … full-batch.
- Biased input distri. … exp(-k) vs $N^{-1/2}$.



Our hypothesis: relative norm adjustments via sampling biases.

- (d, 2)-parity learned with a 2-layer model + correlation loss → track 1 neuron:

$$f(x) = a\sigma(w^\top x) - 1, \quad \sigma(z) := \tfrac{1}{2}z^2.$$

- 2-phase analysis: Full-batch GD (with projection) first on N samples, then population.

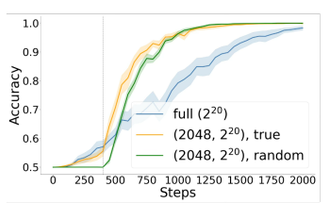
Theorem: #steps for convergence = $\tilde{O}((Nd)^{1/4})$

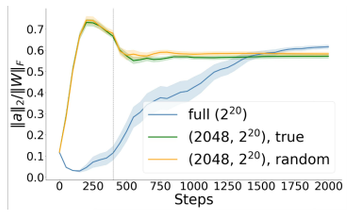
Intuition: the $O(N^{-1/2})$ sampling bias in gradients: smaller N → $|a|/\|w\|$ grows faster
→ larger effective learning rate on w … *interventions with similar effects?*

Empirical evidence via interventions

1. Benefit even from **random-label training**.

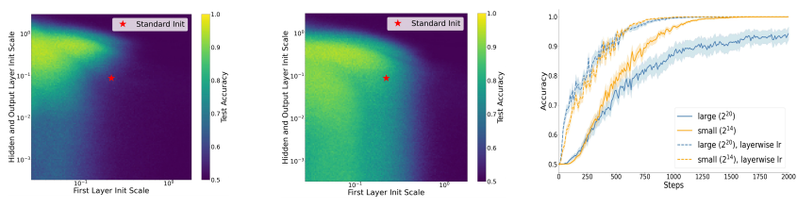
i.e. first train on a small set with random labels, then online training.



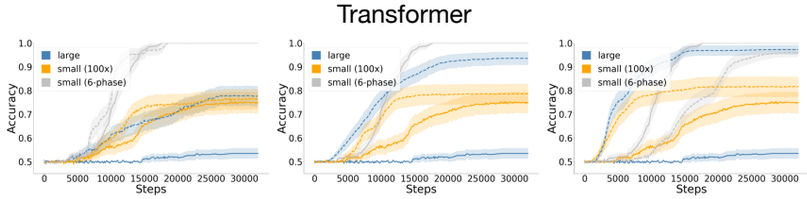


2. **Layerwise interventions**: initialization, learning rates.

MLP



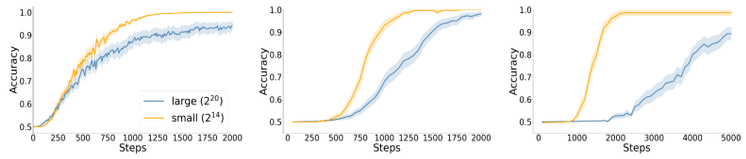
Transformer



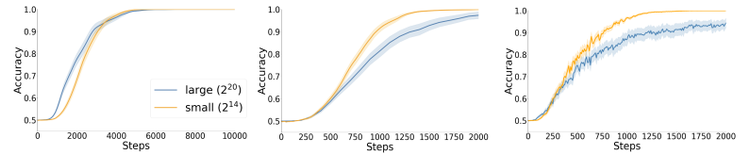
(a) QK learning rate (b) QK initialization scale (c) QK normalization

What makes the small-vs-large gap wider?

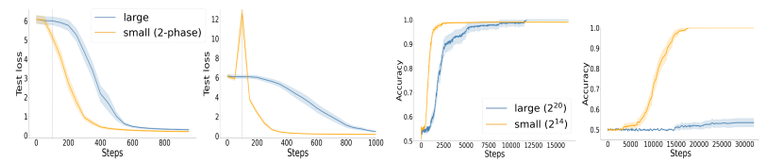
- Models with larger depth (2 → 4 → 6).



- Models with smaller width (1024 → 256 → 64).



- More complex tasks



SIM (dim 40 vs 80) parity ((10,6) vs(20,6))

When is small dataset repetition helpful?

- **Non-convex** optimization (saddle), e.g. deep (linear) nets
- "Hard" tasks with **computational-statistical gaps**.
- **Structured tasks**, e.g. formal reasoning:

  CoT SFT for math (Kopiczko et al. 26).

When is small dataset repetition not helpful?

- Linear regression.
- Dataset has a lot of **redundant data**, e.g. LLM pre-training on web datasets.

Reference

- **Kopiczko et al. 26**: Data Repetition Beats Data Scaling in Long-CoT Supervised Fine-Tuning