On Stronger Computational Separations Between Multimodal and Unimodal Machine Learning

arikarchmer.com

Ari Karchmer
Harvard University
Work done while at Boston University

ArXiv

Multimodal data might not help computationally in practice!

(1) Introduction + Research Questions

- Multimodal data: when many kinds of information are packaged in a single datapoint.
  - For instance, text and image pairs.
- Multimodal machine learning has seen significant empirical success (e.g., GPT-4).
- Finding out **when and why** this success is happening is important.
  (1) Can multimodal data allowing for *computationally faster* learning? More *data efficient* learning?
  (2) What kinds of learning problems are multimodal data useful for?
  (3) Are these learning problems relevant to practice, or very contrived?

★ If we find out the conditions for success, we can more efficiently allocate our resources to take advantage of the benefits of multimodal data in ML.

<u>This paper</u>: investigating (2) and (3) w.r.t. the computational setting, from a theoretical perspective.

Is theory (mathematical formalism) the right perspective ★ for these questions?

Yes—towards ★, we need abstraction. Then, we can use our theory to make predictions about when multimodal data is helpful in ML in practice. This is important in situations where trial and error in performing ML is too expensive.

- "Would harvesting multimodal data solve my compute bottleneck?"
- We should know without needing to harvest data and then check! This could be wasteful if wrong.

(2) Executive Summary

✓ Multimodal machine learning has seen significant empirical success (e.g., GPT-4).
- Zhou Lu (ALT '24) demonstrated a computational separation between multimodal and unimodal learning for **worst-case** instances of a certain learning task.
  - A learning task where multimodal data unlocks otherwise NP-hard learning problems.

✓ **First, this paper presents a stronger <u>*average-case*</u> computational separation, where, for *"typical instances"* of a specific problem:**
- Unimodal learning is computationally hard.
- Multimodal learning is computationally easy.

Hardness holds assuming (worst-case) hardness of Learning Parity with Noise (LPN). Again, for typical instances, not worst-case! This can inform practice **better** than worst-case separations.

✓ **Second, this paper questions the "naturalness" of the average-case learning problem that makes this separation. Would it actually be relevant in practice?**

★<u>Our key finding/theorem</u>★
- Any computational separation between **average-case** unimodal and multimodal learning implies a corresponding **cryptographic key agreement protocol**.

★<u>Interpretation of the theorem</u>★
- Strong computational advantages of multimodal data may occur ***infrequently*** in practice.
- Why? Such advantages exist **only** for "pathological" cases of inherently cryptographic distributions.

★<u>Justification of interpretation</u>★
- A rhetorical question: do you believe that image and text data pairs could be used to encode a cryptographic KA protocol?
- If not, then you buy the interpretation—think about it!

(3) Technical Appetizer

✳ Lu (ALT '24) worst-case model of bimodal and unimodal learning.
- Two "modalities": $X, Y \subseteq \mathbb{R}^n$; label space $Z = \{\pm 1\}$.
- In the bimodal PAC-learning task, selection of a dataset consisting of datapoints abides by a data distribution ρ $X \times Y \times Z$. The goal of a PAC-learning algorithm **A** is to process this dataset to generate a hypothesis function $h : Y \to Z$ that achieves population risk below on the *unimodal* task of labelling elements of Y (w.l.o.g. X).
- $\ell_{\text{pop}}(h) \triangleq \mathbb{E}_{(x,y,z)\sim\rho}[\ell(h, y, z)] \leq \epsilon$ w.p. $1 - \delta$

✳ Average-case bimodal learning (this work).
- Let $\Delta(S)$ denote the convex polytope over all distributions over a set S. We assume that the bimodal learning task is sampled according to a meta-distribution μ over $\Delta(X \times Y \times Z)$. This is a "Bayesian view" of the PAC-learning task, where the learner is assumed to have some prior over the possible data distributions.

✳ Outline of implied cryptographic key agreement.
- Proof sketch: the protocol generates a hard unimodal learning instance by sampling an easy multimodal instance. Security implied because only hard unimodal data exposed to adversary.