On Stronger Computational Separations Between Multimodal and Unimodal Machine Learning

Ari Karchmer [1]

Abstract

Recently, multimodal machine learning has enjoyed huge empirical success (e.g. GPT-4). Motivated to develop theoretical justification for this empirical success, Lu (NeurIPS '23, ALT '24) introduces a theory of multimodal learning, and considers possible *separations* between theoretical models of multimodal and unimodal learning. In particular, Lu (ALT '24) shows a computational separation, which is relevant to *worst-case* instances of the learning task. In this paper, we give a stronger *average-case* computational separation, where for "typical" instances of the learning task, unimodal learning is computationally hard, but multimodal learning is easy. We then question how "natural" the average-case separation is. Would it be encountered in practice? To this end, we prove that under basic conditions, any given computational separation between average-case unimodal and multimodal learning tasks implies a corresponding cryptographic key agreement protocol. We suggest to interpret this as evidence that very strong *computational* advantages of multimodal learning may arise *infrequently* in practice, since they exist only for the "pathological" case of inherently cryptographic distributions. However, this does not apply to possible (super-polynomial) *statistical* advantages.

1. Introduction

For humans, multimodal perception—the ability to interpret the same or similar information expressed in multiple ways (e.g. text and image)—is absolutely critical to learning. We hold it as self-evident that access to multiple representations of the same idea can ease the process of forming a mental model applicable to new situations ("when you put it that way...").

Empirical triumphs of *Machine Learning* from multimodal data such as GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2023), and Gato (Reed et al., 2022) suggest that multimodal perception is also really useful for some machine learning tasks. Lu (2023b;a) introduces a formal study of multimodal versus unimodal machine learning tasks, in order to develop theoretical justification for the empirical results (see also Huang et al. (2021) and others; we elaborate on related work in Section 1.2). However, the theory of multimodal learning is still in its infancy. The main theoretical question is:

> *Is multimodal data truly (provably) more useful than unimodal data, or is it a mirage?*

To attack this question, Lu (2023b) first shows a *statistical* separation: that there exist machine learning tasks that do require asymptotically more samples to complete when the data is expressed unimodally as opposed to multimodally. Second, Lu (2023a) shows that not only is there a statistical separation, but there also exist machine learning tasks that might be *computationally* easier when given access to bimodal data (two modes), rather than just unimodal data. The computational separation of Lu (2023a) identifies a machine learning task that is possible in polynomial time with bimodal data, but not with unimodal data, *for it's worst-case instance*. This means that the unimodal learning task could still possibly be easy on most or "typical" instances. Of course, Lu's separation requires a relatively weak assumption of computational hardness: that a certain NP-hard problem is not also in P.

In this work, we continue to develop a theory of multimodal learning, in pursuit of the truth about how useful multimodal data is (when compared to unimodal data). In particular, we study the existence of stronger computational separations, which apply to the *average-case* instances of learning tasks. An average-case computational separation can inform the practice of multimodal learning more comprehensively than a worst-case separation, since it would apply with high probability over a randomized process that determines the learning task.

[1]Department of Computer Science, Boston University, Boston, MA, USA. Correspondence to: Ari Karchmer <arika@bu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

More specifically (though still informally), an average-case computational separation is a multimodal learning task, where a "typical" instance of the task is learnable in polynomial time, while the corresponding "typical" instance of a *unimodal* learning task is unlearnable in polynomial time. The notion of a typical instance is formalized by considering a fixed *distribution* over learning tasks (and thus some small probability of failure to learn), instead of a universal quantification. We define average-case multimodal learning tasks and computational separations precisely in Section 2.

Our first result is an average-case computational separation, under the computational assumption that learning parities in the presence of a little random learning noise is hard. More specifically,

Definition 1.1 (LPN assumption)**.** For any length parameter $n \in \mathbb{N}$ and noise rate $\theta \in (0, 0.5)$, the $t-\mathrm{LPN}_{\theta,n}$ assumption is that for every probabilistic algorithm $\mathbf{I}$ running in time $t(n)$,

$$\Pr_{\mathbf{x},\mathbf{A},\mathbf{b}}[\mathbf{I}(\mathbf{A}, \mathbf{x}\mathbf{A} + \mathbf{b}) = \mathbf{x}] < 1/t(n)$$

Here, $\mathbf{x}$ is a uniformly random element of $\mathbb{Z}_2^{1\times n}$, $\mathbf{A}$ is a uniformly random element of $\mathbb{Z}_2^{n\times t(n)}$ and $\mathbf{b} \in \mathbb{Z}_2^{1\times t(n)}$ is sampled element-wise from $\mathrm{Ber}(\theta)$.

We construct an average-case computational separation under the $\mathrm{poly}-\mathrm{LPN}_{\theta,n}$ assumption where $\theta \triangleq n^{-0.5}$. We refer informally to this assumption as low-noise LPN.[1]

Theorem 1.2 (Informal)**.** *Under the low-noise LPN assumption, there exists an average-case bimodal learning task that can be completed in polynomial time, and a corresponding average-case unimodal learning task that cannot be completed in polynomial time.*

For simplicity, we prove an average-case computational separation between bimodal and unimodal learning. In the context of a separation, this only strengthens the result, as any separation involving bimodal data applies to the multimodal versus unimodal setting.

Low-noise LPN is a relatively natural hardness of learning assumption. However, the bimodal learning task (and corresponding unimodal task) that we construct to prove the separation is pathologically constructed given the assumption (in Section 1.1 we present a sketch of the idea). Therefore, it makes sense to ask: do there exist more natural bimodal learning tasks that constitute average-case computational separations? Indeed, this question was left open by (Lu, 2023a) even in the context of worst-case separations.

Towards an answer, we look to find the *minimal* computational hardness needed to construct an average-case computational separation between bimodal and unimodal learning. In doing so, we hope to reveal the core computational problem at the center of a separation between multimodal and unimodal learning, so that we can then understand whether it might be frequently encountered in practice.

In this vein, our second result says that to obtain *any* average-case computational separation, we *must* assume enough computational hardness to construct *cryptographic key agreement* protocols. In fact, we construct an explicit key agreement protocol based on any given hypothesized average-case computational separation.[2]

Theorem 1.3 (Informal)**.** *For any given average-case bimodal learning task that can be completed in polynomial time, such that the corresponding unimodal task cannot be completed in polynomial time, there exists a corresponding cryptographic key agreement protocol.*

Cryptographic key agreement (KA)—where two parties communicate over an authenticated but insecure channel in order to jointly agree on a *secret* key—is one of the fundamental tasks of cryptography.[3] The existence of KA protocols is known to be equivalent to the existence of public key encryption schemes, and other exotic cryptographic primitives (see e.g. (Impagliazzo, 1995) for more information).

Interpreting Theorem 1.3. To apply Theorem 1.3 to the question of whether their exist more natural average-case computational separations, we suggest the following perspective. Although Theorem 1.2 gives good evidence that super-polymomial average-case computational separations do *exist*, Theorem 1.3 shows that *any* such separation may not be very "natural," since it needs to be sufficiently "cryptographic." That is, it can be directly used to construct exotic cryptographic primitives. Arguably, "cryptographic" data distributions rarely come up in practice, where data is generated from natural processes instead of the precise design of a cryptographer. Hence, we suggest that **super-polynomial** *computational* advantages of multimodal learning may arise infrequently in practice. Our interpretation seems to contradict the results of practical studies, however our interpretation does not apply to *statistical* advantages of multimodal learning (i.e., less data needed). In fact, even the separation from Theorem 1.2 does not hold in the statistical regime, since LPN can be solved in $2^{o(n)}$ time even with $\mathrm{poly}(n)$

[1] The low-noise LPN assumption is a popular conjecture in the cryptographic literature, as it is known to imply public key encryption (Alekhnovich, 2003), and pseudo-random functions with extremely low circuit depth (Yu and Steinberger, 2016). In particular, it is common to conjecture subexponential (i.e., 2^{n^ϵ}) hardness.

[2] This is a so-called "win-win" result, which may be of independent interest: either secure key agreement protocols exist, or typical instances of unimodal learning tasks can be learned without significantly more computation than the multimodal task.

[3] For example, KA is fundamental to the Transport Layer Security (TLS) protocol for facilitating secure communication over the internet.

samples (Lyubashevsky, 2005). This would explain why multimodal learning continues to succeed in practice: the advantages are typically statistical, not computational.

On polynomial separations. Since Theorem 1.3 derives a cryptographic KA protocol from any given *super*-polynomial separation, we suggest that super-polynomial separations are infrequent in practice. However, polynomial separations (e.g. quadratic computational advantage) may still be relevant in certain practical settings despite not being as totally debilitating as super-polynomial separations. Indeed, a polynomial separation does not necessarily imply (by Theorem 1.3) a KA protocol with cryptographic security (i.e., security against all polynomial time adversaries). Therefore, we (conservatively) refrain from suggesting that polynomial separations are unlikely to occur in practice. That being said, **our proof is general enough to show that any given polynomial separation does still imply a corresponding polynomial-security KA protocol.**[4] Therefore, a more aggressive interpretation could argue that even large polynomial separations (e.g. a quartic computational advantage) are pretty unlikely to occur in practice.

On low-noise LPN. Theorem 1.3 also provides a hint for why we do not achieve a separation as in Theorem 1.2 by using the weaker *standard* LPN assumption, where the noise rate θ can be taken to be any constant fraction less than one half, and the secret parity is uniformly random. Indeed, the standard LPN assumption is not known to imply any form of KA, unlike the low-noise variant. Constructing KA from the standard LPN assumption is a major open problem in the theory of cryptography.

1.1. Our Constructions: the Main Ideas

Theorem 1.2. Loosely speaking, the main idea behind the construction of our average-case computational separation is to use the LPN assumption to obtain a very strong *heterogeneity* between the available data modalities. Heterogeneity is a notion concerning multimodal data studied by (Lu, 2023a), which he identifies as a fundamental aspect of multimodal learning.

Intuitively, the heterogeneity property (in bimodal learning) is that the two modalities somehow complement each other, so that seeing data from both modalities is significantly better than from just one. Indeed, if the two modalities are very similar, then adding data from the second modality is redundant (as an extreme case, consider when datapoints are sampled identically across all modalities).

To construct a *computational* separation, it is clear that the useful second modality should contain information that is *hard to compute* given information in the first modality. Thus, our main idea is to use the trapdoor properties of the LPN assumption to construct a distribution over a modality $\mathcal{Y} \subseteq \mathbb{R}^n$, such that samples from that distribution hide all information about the corresponding sample from a modality $\mathcal{X} \subseteq \mathbb{R}^n$ (with respect to efficient computation). In cryptographic terms, the mapping from $\mathcal{X}$ to $\mathcal{Y}$ is *distributionally one-way*. Since the mapping is one-way, there still exists a learnable *connection* (a mapping) from $\mathcal{Y}$ to $\mathcal{X}$. This one-way connection is an instance of the *connection* property identified by Lu (2023a) as essential to the existence of an advantage from learning with multimodal data.

For our separation, we define the joint data distribution over the two modalities so that the first modality consists of a low-noise LPN instance, and the useful second modality is the parity function that underlies the LPN instance. Obviously, given the LPN assumption, the useful second modality is hard to compute given samples from first. This gives a strong heterogeneity property in a formally justified way.

Simultaneously, we must define the joint distribution over the two modalities so that they can be used to actually learn from the labelled data. Our method to handle this involves injecting hidden data into the first modality which are only recovered given the second modality. Our method is inspired by the ideas behind the low-noise LPN-based public key encryption schemes of (Alekhnovich, 2003), and uses key ideas from the Covert Learning algorithm for noisy parities of (Canetti and Karchmer, 2021).

Theorem 1.3. We have described how an average-case computational separation needs some form of *distributional one-wayness* between the two modalities. This is required because otherwise either modality could be efficiently sampled given the other, making a reduction from unimodal to bimodal learning feasible.

Distributional one-wayness is a standard notion in cryptography, equivalent to the more fundamental notion of one-wayness (Impagliazzo and Levin, 1990). In proving Theorem 1.3, we show that cryptographic key agreement, a cryptographic primitive thought to be (much) stronger than one-way functions, must be an essential aspect of constructing an average-case computational separation.

Our construction of KA exploits the fact that data sampled from the unimodal task is hard to learn from, unless one also has the the corresponding data from the second modality. Towards KA, it suffices to construct a *bit agreement* protocol, before invoking standard techniques from cryptography to obtain a KA protocol for long keys (see Section A for more information). The first player in the bit agreement protocol (called "Alice") uses the average-case computational separation to sample unlabeled bimodal data, and then send only the unimodal data to the second player (called "Bob"). Bob picks a uniformly random bit b_B (which is the bit he

[4] See the proof of Theorem 4.3 for details.

wants Alice to agree with), and if $b_B = 1$, labels the data by a concept sampled according to the multimodal learning task and sends it back, and if $b_B = 0$ responds with uniformly random labels. Alice, given the data labels, applies the multimodal learning algorithm to obtain a hypothesis function. Roughly speaking, Alice can decode Bob's bit with good probability because the accuracy of the hypothesis function resulting from her execution of the multimodal learning algorithm is only good when $b_B = 1$. Finally, the protocol can be shown secure because any polynomial time adversary who, given a view of the protocol, can predict b_B with any probability significantly better than 1/2, can be used to contradict the assumption that the unimodal learning task is hard in polynomial time.

Our bit agreement protocol uses similar ideas to the protocol of Pietrzak and Sjödin (2008), who use the existence of a so-called secret-coin weak pseudorandom function to construct KA. However, our analysis is significantly more involved than that of Pietrzak and Sjödin (2008) since we need to show that an adversary for the protocol implies a unimodal learning algorithm rather than a successful weak pseudorandom function adversary, which is weaker.

1.2. Related Work

Few theoretical results on multimodal learning are known at the moment. For those that exist, they consider certain limited scenarios. For example, works of Yuhas et al. (1989); Sridharan and Kakade (2008); Amini et al. (2009); Federici et al. (2020) consider a situation of learning from multimodal data, but where learning is still possible from each mode individually. This is the so-called *multi-view* setting, and does not produce theoretical justification for any computational separation afforded by access to multimodal data.

On another note, (Huang et al., 2021) study advantages in generalization when learning common latent representations of multimodal data, but not predictors. Additionally, other works like Yang et al. (2015) and Sun et al. (2020) make strong distributional assumptions about the data that is sampled from multiple modalities. It is not clear whether those assumptions hold in practice.

Empirically, deep learning from multimodal data has had great success, for example in learning massive general agents like GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2023) and Gato (Reed et al., 2022). Also, deep learning for modality *generation* has worked well (e.g. Reed et al. (2016) for text to image). It is often observed that ML models derived from multimodal data perform better than even fine-tuned models derived from unimodal data.

2. Technical Overview

Before proving Theorem 1.2 and 1.3 in the next sections, we begin by introducing the model for (average-case) multimodal learning, as well as the notions of computational separations.

2.1. Bimodal Learning

We follow the model for bimodal learning of (Lu, 2023a). In a formal bimodal learning task, two modalities, denoted by $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$, and a label space $\mathcal{Z}$, form the basis for the selection of datapoints $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. In the bimodal PAC-learning task, selection of a dataset consisting of m datapoints abides by a data distribution ρ over $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. For an accuracy and confidence $\epsilon, \delta \in (0, 1)$, the goal of a PAC-learning algorithm A is to process this dataset generated by ρ so as to generate a hypothesis function $h : \mathcal{Y} \to \mathcal{Z}$ that achieves population risk below ϵ, *on the unimodal task of labelling elements of $\mathcal{Y}$ with labels of $\mathcal{Z}$* (and without loss of generality, labelling elements of $\mathcal{X}$):

$$\ell_{\text{pop}}(h) \triangleq \mathop{\mathbb{E}}_{(x,y,z)\sim\rho}\Big[\ell(h, y, z)\Big] \leq \epsilon$$

with probability at least $1 - \delta$, for some loss function ℓ. For example, $\ell_{0-1}(h, y, z) \triangleq \mathbf{1}[h(y) \neq z]$ (0-1 loss).

In Section 3, we also consider $\ell_0(h, y, z) \triangleq \frac{1}{|z|}|\{i : \mathbf{1}[h(y)_i \neq z_i]\}|$ when z is not a single bit. The algorithm A is considered efficient if it runs in polynomial time in the parameters $1/\epsilon, 1/\delta$ and n.

2.2. Relationship Between Modalities

(Lu, 2023a) defined the bimodal PAC-learning task so that there must exist a (unknown) bijection between $\mathcal{X}$ and $\mathcal{Y}$ defined for any (x, y, z) in the support of ρ. In this work we generalize this so that there only exists a (probabilistic) mapping between x and y—we consider this a more practical assumption since many correspondences arising in bimodal learning in practice do not have bijective (or even functional) relationships. For example, consider that a single caption may have many images that it describes, and a single image may have many captions that describe it. Furthermore, for any caption, the paired image sampled by the data distribution may be chosen at random from the set of possible images defined by the mapping.

Probabilistic mappings. Formally, we represent a unidirectional probabilistic mapping from a set S to a set T by a function $\phi : S \to [0, 1]^{|T|}$, subject to the constraint that the sum over all $s \in S$ of $\phi(s) = 1$. The mapping defined by the function ϕ maps an element s to element $t_i \in T$ with probability $\phi(s)_i$. We write $\phi[s]$ to denote a sample from T according to the distribution $\phi(s)$. Frequently, we will define a probabilistic mapping ϕ by explicitly defining the

distribution $\phi[s]$ for all $s \in S$, as it is conceptually easier to define.

2.3. Unimodal Learning

When $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are clear from the context, we identify a bimodal PAC-learning problem by ρ. Arising from a bimodal PAC-learning problem ρ are unimodal PAC-learning problems $\rho_{\mathcal{X},\mathcal{Z}}$ and $\rho_{\mathcal{Y},\mathcal{Z}}$. Here, $\rho_{\mathcal{X},\mathcal{Z}}$ and $\rho_{\mathcal{Y},\mathcal{Z}}$ denote the distribution ρ over $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ projected to $\mathcal{X} \times \mathcal{Z}$ and $\mathcal{Y} \times \mathcal{Z}$ respectively. In unimodal PAC-learning, the task is defined analagously to the bimodal task: the goal of the learning algorithm A for $\rho_{\mathcal{Y},\mathcal{Z}}$ (w.l.o.g.) is to produce a hypothesis h such that

$$\ell_{\text{pop}}(h) \triangleq \mathop{\mathbb{E}}_{(y,z)\sim\rho_{\mathcal{Y},\mathcal{Z}}}\Big[\ell(h,y)\Big] \leq \epsilon$$

with probability greater than $1-\delta$ for some loss function ℓ.

2.4. Average-case Bimodal Learning

In this work, we will primarily consider an average-case notion of bimodal and unimodal learning. Let $\Delta(S)$ denote the convex polytope over all distributions over a set S. In the average-case notion of bimodal learning, we assume that the bimodal learning task is sampled according to a meta-distribution μ over $\Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$. This is consistent with the "Bayesian view" of the PAC-learning task, where the learner is assumed to have some prior over the possible data distributions. When $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ is clear from the context, an average-case bimodal learning problem is identified with μ.

More specifically, we consider the meta-distribution μ to be of the following natural form. Let χ be a fixed distribution over the first modality $\mathcal{X}$. Let η be a distribution over a set of probabilistic mappings $\phi : \mathcal{X} \to [0,1]^{|\mathcal{Y}|}$ that transform elements of the first modality $\mathcal{X}$ to elements of the second modality $\mathcal{Y}$. Finally, let ζ be a distribution over a set of probabilistic mappings $\psi : \mathcal{Y} \to [0,1]^{|\mathcal{Z}|}$. The meta-distribution μ selects a data distribution ρ by sampling $\phi \sim \eta$ and $\psi \sim \zeta$; the data distribution ρ thus samples a datapoint by sampling $x \sim \chi$, and returning $(x, \phi[x], \psi[\phi[x]])$. We write $\mu = (\chi, \eta, \zeta)$ to be explicit about such average-case bimodal learning tasks.

2.5. Average-case Computational Separations

Let us now formalize what is an average-case computational separation in multimodal learning.

Definition 2.1. We say that an average-case bimodal learning task μ (with respect to a loss function ℓ) is a super-polynomial *computational separation* if it holds that:

- There exists a polynomial $p : \mathbb{N} \to \mathbb{N}$ such that there is a time $p(n)$ probabilistic algorithm A such that, when $\rho \sim \mu$, and given access to $p(n)$ datapoints sampled according to ρ, A outputs a hypothesis that achieves population risk $\ell_{\text{pop}}(h) \leq 1/2 - 1/p(n)$ with probability $1/p(n)$ over μ, ρ and randomness of A.
- For $\rho_{\text{uni}} \in \{\rho_{\mathcal{X},\mathcal{Z}}, \rho_{\mathcal{Y},\mathcal{Z}}\}$: For every polynomial $t : \mathbb{N} \to \mathbb{N}$, and every probabilistic algorithm A running in time $t(n)$, when $\rho \sim \mu$, and A is given access to $t(n)$ datapoints sampled according to ρ_{uni}, A outputs a hypothesis such that $\ell_{\text{pop}}(h) > 1/2 - 1/t(n)$ in the unimodal task ρ_{uni} with probability at least $1 - 1/t(n)$ over μ, ρ_{uni} and randomness of A.

We note that the separation only requires population risk $\ell_{\text{pop}}(h) \leq 1/2 - 1/p(n)$ for the bimodal case, and $\ell_{\text{pop}}(h) > 1/2 - 1/t(n)$ for the unimodal case (with high probability). Again, this makes our construction of KA a *stronger* result.

On the other hand, when we construct the separation in section 3, we get a difference in population risk that is optimally large. The learning algorithm for the bimodal task achieves $\ell_{\text{pop}}(h) \leq n^{-0.5}$ (which is optimal), while we prove hardness of achieving $\ell_{\text{pop}}(h) < 1/2 - 1/t(n)$ for any polynomial $t : \mathbb{N} \to \mathbb{N}$ in the unimodal task.

2.6. Relationship to LUPI

The model for multimodal PAC learning defined by Lu (2023a), which is used in this paper, bears resemblance to the Learning Using Privileged Information (LUPI) paradigm of Vapnik and Vashist (2009). In fact, for the bimodal case, the two models of learning are the same (we omit formal proof, which follows immediately by definition).

Both LUPI and bimodal learning consider triplets of information (rather than the standard of pairs). These triplets are given as input to the learning algorithm at training time, while at test time, only pairs are received. In this way, the two models consider situations where multiple modalities are accessible for training machine learning models, and may be effective for learning problems that act on just a single modality at test time. In the case of LUPI, the second modality is motivated by the presence of a "teacher." The "teacher" can try to ease the learning process by giving the learner some auxiliary information about the data. In the case of multimodal learning, the additional modalities are motivated by the contemporary success of multimodal perception in AI, where there is an abundance of data but not necessarily a "teacher."

Since the LUPI paradigm of Vapnik and Vashist (2009) is the same as the bimodal learning model of Lu (2023a), both our main results (Theorem 1.2 and 1.3) apply to the LUPI paradigm. Previous work (Vapnik and Vashist, 2009; Lapin et al., 2014; Vapnik et al., 2015) on understanding the LUPI

paradigm (e.g. proving upper bounds, lower bounds, and separations) focused on statistical learning settings such as empirical risk minimization (ERM). To our knowledge, our results are the first to consider the average-case computational power of the LUPI paradigm.

3. Average-Case Separation

In this section, we construct a super-polynomial computational separation, assuming hardness of the $t - \mathrm{LPN}_{n^{-0.5},n}$ problem. We recall that an average-case multimodal learning problem $\mu = (\chi, \eta, \zeta)$ consists of:

- χ: a distribution over the modality $\mathcal{X}$.
- η: a distribution over probabilistic mappings that transform $\mathcal{X} \to \mathcal{Y}$.
- ζ: a distribution over probabilistic mappings that transform $\mathcal{Y} \to \mathcal{Z}$.

In order to construct our separation, we will only need to define η so that it places the entire probability mass on a single probabilistic mapping $\phi : \mathcal{X} \to [0,1]^{|\mathcal{Y}|}$. However, because we prove a separation (i.e., a "negative" result), this only strengthens the result.

3.1. Construction of Separation

Let $\mathrm{Ber}(m)$ denote the Bernoulli random variable with mean $m \in [0,1]$.

Consider the following average-case multimodal learning problem $\mu = (\chi, \eta, \zeta)$. The modalities are $\mathcal{X} = \mathbb{Z}_2^{1\times n} \times [n]$, $\mathcal{Y} = \mathbb{Z}_2^{n\times n} \times \mathbb{Z}_2^{1\times n}$, and $\mathcal{Z} = \mathbb{Z}_2^{n\times 1} \times \mathbb{Z}_2$. All sums are computed modulo 2.

- χ: Sample uniformly random $i \in [n]$, and $\mathbf{x} \in \mathbb{Z}_2^{n\times 1}$ where $\mathbf{x}_j$ is sampled i.i.d. from $\mathrm{Ber}(n^{-0.5})$. Output $(\mathbf{x}, i)$.
- η: With probability 1, output the probabilistic mapping $\phi : \mathcal{X} \to [0,1]^{|\mathcal{Y}|}$. We define

$$\phi[(\mathbf{x}, i)] = (\mathbf{A}, \mathbf{x}\mathbf{A} + \mathbf{b} + \mathbf{e}^{(i)})$$

  where $\mathbf{A} \in \mathbb{Z}_2^{n\times n}$ is a uniformly random, $\mathbf{b} \in \mathbb{Z}_2^{1\times n}$ is such that $\mathbf{b}_i$ is sampled i.i.d. from $\mathrm{Ber}(n^{-0.5})$, and $\mathbf{e}^{(i)} \in \mathbb{Z}_2^{1\times n}$ is defined so that $(\mathbf{e}^{(i)})_j = 1$ if and only if $j = i$.
- ζ: Sample probabilistic mapping $\psi_{\mathbf{w}} : \mathcal{Y} \to [0,1]^{|\mathcal{Z}|}$ by sampling random vector $\mathbf{w} \in \mathbb{Z}_2^{n\times 1}$ such that $\mathbf{w}_i$ is sampled i.i.d. from $\mathrm{Ber}(n^{-0.5})$. We define

$$\psi_{\mathbf{w}}[(\mathbf{Y}, \mathbf{y})] = (\mathbf{Y}\mathbf{w} + \mathbf{b}', \mathbf{y}\mathbf{w} + \mathbf{b}'')$$

  where $\mathbf{b}' \in \mathbb{Z}_2^{n\times 1}$ is such that $\mathbf{b}'_i$ is sampled i.i.d. from $\mathrm{Ber}(n^{-0.5})$. The bit $\mathbf{b}''$ is also sampled i.i.d. from $\mathrm{Ber}(n^{-0.5})$.

Theorem 3.1 (Separation)**.** *Under the* $\mathrm{poly}-\mathrm{LPN}_{\theta,n}$ *assumption for for* $\theta \triangleq n^{-0.5}$*, the multimodal learning task* $\mu = (\chi, \eta, \zeta)$*, as defined above, is a super-polynomial computational separation.*

Proof. The statement follows immediately from Theorem 3.2 and Theorem 3.4. □

We prove Theorems 3.2 and 3.4 in the following two sections.

3.2. An Efficient Multimodal Learning Algorithm

We begin by proving that there exists an efficient algorithm for the multimodal PAC-learning task defined by μ. This is the part of the separation that shows the feasibility of the learning task given bimodal data.

Now, observe that, given samples of the form $(x, y, z) \sim \rho$ for $\rho \sim \mu$, the multimodal PAC-learning task is learned optimally by finding the vector $\mathbf{w}$ underlying $\psi_{\mathbf{w}}$. Hence, we now give an algorithm that finds $\mathbf{w}$, and then outputs the optimal hypothesis.

Algorithm 1 $A_\mu \mid \rho \sim \mu$

Input: n^3 samples $(x, y, z) \sim \rho$.
Output: $h : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$.
Interpret each example (x_j, y_j, z_j) as $((\mathbf{x}_j, i_j), (\mathbf{Y}_j, \mathbf{y}_j), (\mathbf{z}_j, z_j))$.
Sort all examples $((\mathbf{x}_j, i_j), (\mathbf{Y}_j, \mathbf{y}_j), (\mathbf{z}_j, z_j))$ into n bins labelled by $i \in [n]$ by value of i_j.
for each bin b_i **do**
  for example $((\mathbf{x}_j, i_j), (\mathbf{Y}_j, \mathbf{y}_j), (\mathbf{z}_j, z_j))$ in bin b_i: **do**
    Compute $\alpha_{i,j} = \mathbf{x}_j\mathbf{z}_j + z_j$
  end for
  Compute $\mathbf{w}'_i$ by taking the majority vote over $\alpha_{i,j}$ for all j.
end for
Output $h((\mathbf{Y}, \mathbf{y})) = (\mathbf{Y}\mathbf{w}', \mathbf{y}(\mathbf{w}')^{\mathsf{T}})$.

We prove that with high probability, the algorithm A_μ outputs $\mathbf{w}' = \mathbf{w}$, and this minimizes population risk for the PAC-learning task, with respect to ℓ_0 loss.

Theorem 3.2. *We have that*

$$\Pr_{A_\mu, \rho\sim\mu}[\ell_{\mathrm{pop}}(h) \le n^{-0.5} : h \leftarrow A_\mu] \ge 1 - \exp(-\Omega(n))$$

with respect to ℓ_0 *loss. Moreover,* A_μ *runs in time* $\mathrm{poly}(n)$.

Proof. The runtime of A_μ being $\text{poly}(n)$ is immediate.

To show that population risk is small, we need to show that:

$$\Pr_{A_\mu,\rho\sim\mu}\left[\mathop{\mathbb{E}}_{(x,y,z)\sim\rho}\left[\ell_0(h,y)\right] \le 1/n^{0.5} : h \leftarrow A_\mu\right] \ge 1-\exp(-\Omega(n))$$

Consider that if A_μ outputs hypothesis h such that $\mathbf{w}' = \mathbf{w}$, where $\mathbf{w}$ is the vector sampled by ζ, then h satisfies

$$\mathop{\mathbb{E}}_{(x,y,z)\sim\rho}\left[\ell_0(h,y)\right] \le 1/n^{0.5}$$

Thus we will prove that A_μ finds $\mathbf{w}' = \mathbf{w}$ with probability at least $1-\exp(-\Omega(n))$ over the randomness of $\rho\sim\mu$ and the n^3 examples sampled from ρ given as input to A_μ.

To prove this, let us focus on bit $\mathbf{w}'_i$, without loss of generality (the following argument is applies to all $i\in[n]$). The bit $\mathbf{w}'_i$ is the majority vote of $\alpha_{ij} = \mathbf{x}_j\mathbf{z}_j + z_j$ for all examples $((\mathbf{x}_j,i_j),(\mathbf{Y}_j,\mathbf{y}_j),(\mathbf{z}_j,z_j))$ conditioned on $i_j = i$. Therefore, if

$$\Pr_{((\mathbf{x}_j,i_j),(\mathbf{Y}_j,\mathbf{y}_j),(\mathbf{z}_j,z_j))}[\mathbf{x}_j\mathbf{z}_j + z_j = \mathbf{w}_i | i_j = i] \quad (1)$$

$$\ge 1/2 + \Omega(1) \quad (2)$$

then $(\mathbf{w}')_i = \mathbf{w}_i$ with probability at least $1-\exp(-\Omega(n))$, given enough voter participation. Leaving aside the issue of number of votes, note that, by a union bound it would follow that $\mathbf{w}' = \mathbf{w}$ still with probability $1-\exp(-\Omega(n))$ as desired. We now show (1), and leave the issue of lower bounding the number of votes (i.e., the number of examples in every bucket) for after.

To show (1), we expand:

$$\begin{aligned}\mathbf{x}_j\mathbf{z}_j + z_j &= \mathbf{x}_j(\mathbf{A}_j\mathbf{w}+\mathbf{b}') + (\mathbf{x}_j\mathbf{A}_j + \mathbf{b} + \mathbf{e}^{(i)})\mathbf{w} + \mathbf{b}'' \\ &= \mathbf{x}_j(\mathbf{A}_j\mathbf{w}) + \mathbf{x}_j\mathbf{b}' + \mathbf{x}_j(\mathbf{A}_j\mathbf{w}) + \mathbf{b}\mathbf{w} \\ &\quad + \mathbf{e}^{(i)}\mathbf{w} + \mathbf{b}'' \\ &= \mathbf{x}_j\mathbf{b}' + \mathbf{b}\mathbf{w} + \mathbf{e}^{(i)}\mathbf{w} + \mathbf{b}''\end{aligned}$$

And now we argue that for $n > 4$,

$$\Pr\left[\mathbf{x}_j\mathbf{b}' + \mathbf{b}\mathbf{w} + \mathbf{e}^{(i)}\mathbf{w} + \mathbf{b}'' = \mathbf{w}_i\right] \ge 0.515$$

To see this, observe that $\mathbf{e}^{(i)}\mathbf{w} = \mathbf{w}_i$, so it suffices to show

$$\Pr\left[\mathbf{x}_j\mathbf{b}' + \mathbf{b}\mathbf{w} + \mathbf{b}'' = 0\right] \ge 0.515$$

Each term forming the sum inside the probability is an independent random variable. Thus, let us lower bound the probability that each of the three terms is 0. For the first term, we have that $\mathbf{x}_j\mathbf{b}' = \sum i^n(\mathbf{x}_j)_i(\mathbf{b}')_i$ and for each i, $\Pr[(\mathbf{x}_j)_i(\mathbf{b}')_i = 1] = 1/n$ (by definition of the sampling process, where $(\mathbf{x}_j)_i$ and $(\mathbf{b}')_i$ are 1 with probability $1/n^{0.5}$). Hence, for $n > 4$, for all i, $(\mathbf{x}_j)_i$ and $(\mathbf{b}')_i$ are = 0, with probability at least 0.326 (direct computation). Therefore, $\Pr[\mathbf{x}_j\mathbf{b}' = 0] \ge 0.5 + 0.326/2 \ge 0.663$.

The same argument and conclusion holds for the second term. For the third term, we know that $\Pr[\mathbf{b}'' = 1] = 1/n^{0.5}$. Therefore,

$$\Pr\left[\mathbf{x}_j\mathbf{b}' = \mathbf{b}\mathbf{w} = \mathbf{b}'' = 0\right] \ge 0.663^2 \cdot 0.9 \ge 0.395$$

Also,

$$\Pr\left[\mathbf{x}_j\mathbf{b}' = \mathbf{b}\mathbf{w} = 1 \wedge \mathbf{b}'' = 0\right] \ge 1/e^2 \cdot 0.9 \ge 0.12$$

So we conclude that

$$\Pr\left[\mathbf{x}_j\mathbf{b}' + \mathbf{b}\mathbf{w} + \mathbf{b}'' = 0\right] \ge 0.515$$

Thus, if the number of examples in each bin b_i is at least n, then by standard application Chernoff bounds (see Lemma 3.3), the majority vote over all $\alpha_{i,j}$ used to compute $\mathbf{w}'_i$ matches $\mathbf{w}_i$, save for an event of exponentially small probability in n. Furthermore, another application of Chernoff and union bounds gives that the number of examples in every bin is at least n, save for a bad event that occurs with exponentially small probability in n. A final union bound concludes that A_μ finds $\mathbf{w}' = \mathbf{w}$ with probability at least $1-\exp(-\Omega(n))$ as desired. □

Lemma 3.3 (Chernoff Bound, cf. Theorem 2.1 (Janson et al., 2011))**.** *Let $X \sim \text{Bin}(m,p)$ and $\lambda = m\cdot p$. For any $t \ge 0$,*

$$\Pr[|X - \mathbb{E}\left[X\right]| \ge t] \le \exp\left(\frac{-t^2}{2(\lambda + t/3)}\right)$$

3.3. Hardness for Unimodal Learning

Now that we have shown that the bimodal learning problem is learnable in polynomial time, we will show that the corresponding unimodal task, cannot be learned in polynomial time, unless the low-noise LPN assumption does not hold with respect to polynomial time adversaries.

Theorem 3.4. *Let $\mu = (\chi,\eta,\zeta)$ be defined as in section 3.1. Assume that the $\text{poly}-\text{LPN}_{\theta,n}$ assumption holds for $\theta \triangleq n^{-0.5}$. Then, for every polynomial $t:\mathbb{N}\to\mathbb{N}$, and every probabilistic algorithm A running in time $t(n)$, when $\rho\sim\mu$, and A is given access to $t(n)$ datapoints sampled according to $\rho_{\text{uni}} \in \{\rho_{\mathcal{X},\mathcal{Z}}, \rho_{\mathcal{Y},\mathcal{Z}}\}$, A outputs a hypothesis such that $\ell_{\text{pop}}(h) > 1/2 - 1/t(n)$ in the unimodal task with probability at least $1-1/t(n)$ over μ, ρ and randomness of A.*

To prove the theorem, we exploit the *decisional* version of the LPN assumption. Informally, the decisional LPN assumption is that for a suitably defined distribution of

LPN samples, it is hard to distinguish them from uniformly random bits. Most importantly, the decisional LPN and *search* LPN (definition 1.1) are equivalent for the $\text{poly}-\text{LPN}_{n^{-0.5},n}$ regime, due to the existence of a polynomial time serach-to-decision reduction (consult Pietrzak (2012) for more information, and Lemma 1 of Katz et al. (2010) for the search-to-decision reduction).

Furthermore, Applebaum et al. (2009) introduce an important variant of the problem, where the secret parity function is sampled from the same distribution as the noise vector. Applebaum et al. (2009) show that this variant is as hard as when the secret parity function is sampled uniformly at random. We state this variant of the decisional assumption below.

Definition 3.5 (Decisional LPN)**.** For any length parameter $n \in \mathbb{N}$ and noise rate $\theta \in (0, 0.5)$, the $t-\text{DLPN}_{\theta,n}$ assumption is that for every probabilistic algorithm $\mathbf{D}$ running in time $t(n)$,

$$\left| \Pr_{\mathbf{D},\mathbf{A},\mathbf{x},\mathbf{b}} [\mathbf{D}(\mathbf{A}, \mathbf{xA} + \mathbf{b}) = 1] - \Pr_{\mathbf{D},\mathbf{u}} [\mathbf{D}(\mathbf{A}, \mathbf{u}) = 1] \right| < 1/t(n)$$

Here, $\mathbf{A}$ is a uniformly random element of $\mathbb{Z}_2^{n \times t(n)}$, and $\mathbf{x}, \mathbf{b} \in \mathbb{Z}_2^{1 \times t(n)}$ are sampled element-wise from $\text{Ber}(\theta)$, while $\mathbf{u}$ is a uniformly random element of $\mathbb{Z}_2^{1 \times t(n)}$.

Proof of Theorem 3.4. See appendix section B. □

4. KA from Computational Separations

In this section, we construct a cryptographic key agreement protocol, given a super-polynomial computational separation for a multimodal *binary classification* task μ. Here, we have modalities $\mathcal{X}, \mathcal{Y}$, and the label space $\mathcal{Z}$ is fixed to $\{0, 1\}$. We assume that the computational separation is with respect to ℓ_{0-1} loss, since μ is a multimodal binary classification task.

To construct cryptographic key agreement, it is only necessary to construct a bit agreement protocol. A bit agreement protocol is key agreement for a key of one bit, with a small but nontrivial probability of agreement better than 1/2. By standard cryptographic techniques—parallel repetition and privacy amplification—a bit agreement protocol can be converted in to a full-blow key agreement protocol for long keys. Thus, we will construct a bit agreement protocol here. We refer the reader to (Holenstein, 2006) for more information about constructing key agreement from bit agreement. We give a formal definition of bit agreement in the appendix.

4.1. A Bit Agreement Protocol

Consider the protocol between players Alice and Bob specified below.

Algorithm 2 Protocol 1

1: Alice samples $x_1, \cdots x_{k+1} \sim \chi$ and $\phi \sim \eta$.
2: Alice computes $y_i = \phi[x_i]$ for all $i \in [k+1]$.
3: Alice sends Bob $(y_i)_{i \in [k+1]} \in \mathcal{Y}^{k+1}$.
4: Bob samples a bit $b_B \in \{0, 1\}$.
5: If $b_B = 0$, Bob samples a random string $w \in \mathcal{Z}^{k+1}$, and sends Alice w.
6: If $b_B = 1$, Bob samples $\psi \sim \zeta$, and computes $z_i = \psi[y_i]$ for all $i \in [k+1]$, and sends Alice $(z_i)_{i \in [k+1]} \in \mathcal{Z}^{k+1}$.
7: Alice interprets the first t bits of the string she received by considering the i^{th} bit a label for the multimodal datapoint (x_i, y_i). Alice runs a multimodal learning algorithm, using the datapoints $(x_i, y_i, z_i)_{i \in [k]}$, for $\mu = (\chi, \eta, \zeta)$, to obtain a hypothesis h.
8: Bob outputs b_B. Alice outputs $\mathbf{1}[h(y_{k+1}) = z_{k+1}]$.

Theorem 4.1. *Assume that χ, η and ζ are samplable in time $\text{poly}(n)$. If $\mu = (\chi, \eta, \zeta)$ is an average-case super-polynomial computational separation, then there exist a cryptographic key agreement protocol.*

Proof. The statement follows from theorem 4.2 and 4.3, which prove that protocol 1 is a secure and correct bit agreement protocol, and then applying parallel repetitions and privacy amplification. □

Theorem 4.2 (Correctness of BA protocol)**.** *Suppose that $\mu = (\chi, \eta, \zeta)$ is a super-polynomial computational separation. Also, assume that χ, η and ζ are samplable in time $\text{poly}(n)$. Then, there exists a polynomial $q : \mathbb{N} \to \mathbb{N}$ such that Alice and Bob output the same bit with probability at least $1/2 + 1/q(n)$. In other words,*

$$\mathbb{E}\Big[\mathbf{1}[b_B = \mathbf{1}[h(y_{k+1}) = z_{k+1}]]\Big] \geq 1/2 + 1/q(n)$$

Additionally, Alice and Bob each run in polynomial time.

Proof. Conditioning on $b_B = 0$, we know that Alice outputs 0 with probability exactly $1/2$, since in this case Bob chose z_{k+1} uniformly at random.

Conditioning on $b_B = 1$, we now know that because $\mu = (\chi, \eta, \zeta)$ is a, average-case super-polynomial computational separation, there exists a polynomial p and a time $p(n)$ probabilistic algorithm A such that, when $\rho \sim \mu$, and given access to datapoints sampled according to ρ (the bimodal data!), A outputs a hypothesis $\ell_{\text{pop}}(h) \leq 1/2 - p(n)$ with probability $1/p(n)$ over μ, ρ and randomness of A. Thus, when Alice runs this bimodal learning algorithm on the

dataset $(x_i, y_i, z_i)_{i\in[k]}$, she obtains a hypothesis h such that $\ell_{\text{pop}}(h) \leq 1/2 - p(n)$ with probability at at least $1/p(n)$, and with remaining probability has ℓ_{pop} at worst equal to $1/2 + \nu(n)$. If ℓ_{pop} was larger than 1/2 + any negligible function of n, then her hypothesis could be efficiently tested and negated to obtain one with $\ell_{\text{pop}}(h) \leq 1/2 - q(n)$.

Therefore, using that $\Pr[b_B = 0] = 1/2$, we can conclude that

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{1}[b_B = \mathbf{1}[h(y_{k+1}) = z_{k+1}]]\Big] \\
&\geq \frac{1}{2}\left(\frac{1/2 + 1/p(n)}{p(n)} + \left(\frac{1}{2} - \nu(n)\right)\left(1 - \frac{1}{p(n)}\right)\right) \\
&\quad + \left(\frac{1}{2}\right)^2 \\
&\geq \frac{1}{2} + \frac{1}{p(n)^2} - \nu(n)
\end{aligned}$$

Finally, it is immediate that both Alice and Bob run in polynomial time, since χ, η and ζ are polynomial time samplable, and due to the fact that by assumption there exists a polynomial time bimodal learning algorithm for μ. This suffices to complete the proof of the theorem. □

We now prove security of the protocol; that is, an adversary who views the interaction between Alice and Bob, denoted by $\text{View}(\text{A} \leftrightarrow \text{B})$, can predict the bit output by Bob with probability at most $1/2 + \nu(n)$ for a negligible function ν.

Theorem 4.3 (Security of BA protocol)**.** *Suppose that $\mu = (\chi, \eta, \zeta)$ is a super-polynomial computational separation. Then, for any polynomial t, and algorithm D running in time $t(n)$,*

$$\mathbb{E}\Big[D(\text{View}(\text{A} \leftrightarrow \text{B})) = b_B\Big] < 1/2 + 1/t(n)$$

Proof Sketch. We prove the theorem by applying old techniques from the theory of pseudorandomness and cryptography. In particular, a standard "hybrid argument" (Goldwasser and Micali, 1982) together with a reduction from learning to next-bit prediction in the spirit of (Yao, 1982). See appendix C for a detailed proof. □

5. Conclusion

In addition to the technical results, the key *conceptual* contribution of this work is an heuristic argument that the advantages of multimodal machine learning observed in practice are typically statistical, not computational. That is, multimodal perception typically allows for good training with less data, though perhaps not significantly less computation. Our argument relies on the fact that we can directly and explicitly construct a KA protocol from any given average-case super-polynomial computational separation. Thus, average-case super-polynomial computational separations may not arise naturally in real world data. However, KA does not follow *solely* from any average-case statistical separation, so strong statistical advantages may still be encountered frequently. KA cannot follow from *solely* from any average-case statistical separation, because KA fundamentally requires a computational advantage, while a statistical separation considers only unbounded computational parties and no computational-statistical gaps.

Future work. More work can be done to continue to understand the theoretical foundations of multimodal learning. For example, we show that average-case *super-polynomial* computational separations imply cryptographic KA protocols that have super-polynomial security. We use this to present a heuristic argument that such super-polynomial computationally advantages might be rare in the real world. However, if we only assume a polynomial separation (e.g. quadratically more computation is necessary in unimodal learning), then this would could still be relevant to practice, but fundamentally separate from traditional cryptography (which considers super-polynomial adversaries). **That being said, our proof is general enough to show that any given polynomial separation does still imply a polynomial-security KA protocol.** To study the polynomial regime further, we propose to investigate relationships to *fine-grained* public key cryptography (LaVigne et al., 2019).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This work was supported by the DARPA SIEVE program, Agreement Nos. HR00112020021 and HR00112020023. The author thanks Zhou Lu for several helpful conversations, and Sivan Sabato for pointing out a relationship to the LUPI paradigm.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Michael Alekhnovich. More on average case vs approxi-

mation complexity. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 298–307. IEEE, 2003.

Massih R Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. *Advances in neural information processing systems*, 22, 2009.

Benny Applebaum, David Cash, Chris Peikert, and Amit Sahai. Fast cryptographic primitives and circular-secure encryption based on hard learning problems. In *Annual International Cryptology Conference*, pages 595–618. Springer, 2009.

Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

Ran Canetti and Ari Karchmer. Covert learning: How to learn with an untrusted intermediary. In *Theory of Cryptography: 19th International Conference, TCC 2021, Raleigh, NC, USA, November 8–11, 2021, Proceedings, Part III 19*, pages 1–31. Springer, 2021.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

Shafi Goldwasser and Silvio Micali. Probabilistic encryption & how to play mental poker keeping secret all partial information. In *Proceedings of the fourteenth annual ACM symposium on Theory of computing*, pages 365–377, 1982.

Thomas Holenstein. *Immunization of key-agreement schemes*. PhD thesis, PhD thesis, ETH Zürich, 2006.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multimodal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

Russell Impagliazzo. A personal view of average-case complexity. In *Proceedings of Structure in Complexity Theory. Tenth Annual IEEE Conference*, pages 134–147. IEEE, 1995.

Russell Impagliazzo and LA Levin. No better ways to generate hard np instances than picking uniformly at random. In *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, pages 812–821. IEEE, 1990.

Svante Janson, Tomasz Luczak, and Andrzej Rucinski. *Random graphs*. John Wiley & Sons, 2011.

Ari Karchmer. Distributional pac-learning from nisan's natural proofs. In *15th Innovations in Theoretical Computer Science Conference (ITCS 2024)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2024.

Jonathan Katz, Ji Sun Shin, and Adam Smith. Parallel and concurrent security of the hb and hb+ protocols. *Journal of cryptology*, 23(3):402–421, 2010.

Maksim Lapin, Matthias Hein, and Bernt Schiele. Learning using privileged information: Svm+ and weighted svm. *Neural Networks*, 53:95–108, 2014.

Rio LaVigne, Andrea Lincoln, and Virginia Vassilevska Williams. Public-key cryptography in the fine-grained setting. In *Annual International Cryptology Conference*, pages 605–635. Springer, 2019.

Zhou Lu. On the computational benefit of multimodal learning. *arXiv preprint arXiv:2309.13782*, 2023a.

Zhou Lu. A theory of multimodal learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Vadim Lyubashevsky. The parity problem in the presence of noise, decoding random linear codes, and the subset sum problem. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 378–389. Springer, 2005.

Krzysztof Pietrzak. Cryptography from learning parity with noise. In *International Conference on Current Trends in Theory and Practice of Computer Science*, pages 99–114. Springer, 2012.

Krzysztof Pietrzak and Johan Sjödin. Weak pseudorandom functions in minicrypt. In *Automata, Languages and Programming: 35th International Colloquium, ICALP 2008, Reykjavik, Iceland, July 7-11, 2008, Proceedings, Part II 35*, pages 423–436. Springer, 2008.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. 2008.

Xinwei Sun, Yilun Xu, Peng Cao, Yuqing Kong, Lingjing Hu, Shanghang Zhang, and Yizhou Wang. Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 171–188. Springer, 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Vladimir Vapnik, Rauf Izmailov, et al. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1):2023–2049, 2015.

Yang Yang, Han-Jia Ye, De-Chuan Zhan, and Yuan Jiang. Auxiliary information regularized machine for multiple modality feature learning. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Andrew C Yao. Theory and application of trapdoor functions. In *23rd Annual Symposium on Foundations of Computer Science (SFCS 1982)*, pages 80–91. IEEE, 1982.

Yu Yu and John Steinberger. Pseudorandom functions in almost constant depth from low-noise lpn. In *Advances in Cryptology–EUROCRYPT 2016: 35th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vienna, Austria, May 8-12, 2016, Proceedings, Part II 35*, pages 154–183. Springer, 2016.

Ben P Yuhas, Moise H Goldstein, and Terrence J Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27 (11):65–71, 1989.

A. Bit Agreement

A bit agreement protocol is communication protocol between two parties, Alice and Bob. Alice and Bob are allowed to communicate over an authenticated (but insecure) channel. They each begin with a common input n delivered in unary (1^n) which constitutes a security parameter. At the end of their communication, Alice and Bob each output a single bit, b_A and b_B respectively.

We say that the bit agreement protocol has *correctness* if there exists some polynomial $q : \mathbb{N} \to \mathbb{N}$ such that:

$$\mathbb{E}\Big[\mathbf{1}[b_A = b_B]\Big] \geq 1/2 + \frac{1}{q(n)}$$

We say that the bit agreement protocol is secure if, for any polynomial $p : \mathbb{N} \to \mathbb{N}$, and any probabilistic $p(n)$ time adversary D,

$$\mathbb{E}\Big[D(\mathrm{View}(\mathrm{A} \leftrightarrow \mathrm{B})) = b\Big] < 1/2 - \frac{1}{p(n)}$$

Here $\mathrm{View}(\mathrm{A} \leftrightarrow \mathrm{B})$ is defined as the entire transcript of the communication between Alice and Bob.

Key Agreement from Bit Agreement. A secure and correct bit agreement protocol can be transformed into a full-blow cryptographic key agreement protocol. Roughly speaking, this is done by repeating the protocol several times in parallel, and then applying standard privacy amplification techniques to derive totally hidden and uniformly random keys. We refer to (Holenstein, 2006) for details of privacy amplification.

B. Proof of Theorem 3.4

Theorem B.1 (Theorem 3.4 restated)**.** *Let $\mu = (\chi, \eta, \zeta)$ be defined as in section 3.1. Assume that the $\mathrm{poly}{-}\mathrm{LPN}_{\theta,n}$ assumption holds for $\theta \triangleq n^{-0.5}$. Then, for every polynomial $t : \mathbb{N} \to \mathbb{N}$, and every probabilistic algorithm A running in time $t(n)$, when $\rho \sim \mu$, and A is given access to $t(n)$ datapoints sampled according to $\rho_{\mathrm{uni}} \in \{\rho_{\mathcal{X},\mathcal{Z}}, \rho_{\mathcal{Y},\mathcal{Z}}\}$, A outputs a hypothesis such that $\ell_{\mathrm{pop}}(h) > 1/2 - 1/t(n)$ in the unimodal task with probability at least $1 - 1/t(n)$ over μ, ρ and randomness of A.*

Proof of Theorem 3.4. We may assume $\mathrm{poly}{-}\mathrm{LPN}_{\theta,n}$ for $\theta \triangleq n^{-0.5}$, else the statement is vacuous.

We begin by considering the case of $\rho_{\mathrm{uni}} = \rho_{\mathcal{Y},\mathcal{Z}}$. Towards a contradiction, we will prove that if there exists a polynomial $t : \mathbb{N} \to \mathbb{N}$, such that there exists a probabilistic algorithm A' running in time $t(n)$, when $\rho \sim \mu$, and A' is given access to $t(n)$ datapoints sampled according to $\rho_{\mathcal{Y},\mathcal{Z}}$, A' outputs a hypothesis such that $\ell_{\mathrm{pop}}(h) \leq 1/2 - 1/t(n)$ in the unimodal task with probability at least $1/t(n)$ over μ, ρ and randomness of A', then $\mathrm{poly}{-}\mathrm{DLPN}_{\theta,n}$ for $\theta \triangleq n^{-0.5}$ does not hold. This suffices to contradict $\mathrm{poly}{-}\mathrm{LPN}_{n^{-0.5},n}$, due to the polynomial equivalence between the two (Katz et al., 2010).

More specifically and formally, assume A satisfies

$$\Pr_{A',\rho\sim\mu}\left[\mathop{\mathbb{E}}_{(y,z)\sim\rho_{\mathcal{Y},\mathcal{Z}}}\Big[\ell_0(h,y)\Big] \leq 1/2 - 1/t(n) : h \leftarrow A'\right] \geq 1/t(n) \tag{3}$$

We will show that $\mathrm{poly}{-}\mathrm{DLPN}_{\theta,n}$ for $\theta \triangleq n^{-0.5}$ does not hold, and conclude the statement by invoking the polynomial time reduction to $\mathrm{poly}{-}\mathrm{LPN}_{\theta,n}$.

Suppose that A' uses $m(n)$ examples sampled from $\rho_{\mathcal{Y},\mathcal{Z}}$. Construct a distinguisher $\mathbf{D}$ as follows. Given input $(\mathbf{A}, \mathbf{q})$ of the form $\mathbb{Z}_2^{n\times m(n)n} \times \mathbb{Z}_2^{1\times m(n)n}$, use it to sample $m(n)$ tuples of the form $((\mathbf{Y}, \mathbf{y}), (\mathbf{z}, z))$, as the distribution $\rho_{\mathcal{Y},\mathcal{Z}}$ would (here $\mathbf{Y} \in \mathbb{Z}_2^{n\times n}$ and the rest are defined analogously). This can be done by sampling $\psi_{\mathbf{w}} \sim \zeta$, and computing $(\mathbf{z}_i, z_i) = \psi_{\mathbf{w}}[(\mathbf{Y}_i, \mathbf{y}_i)]$ where $(\mathbf{Y}_i, \mathbf{y}_i)$ is the i^{th} contiguous block of length n sliced out from $(\mathbf{A}, \mathbf{q})$, and $\mathbf{y}_i$ has a random bit negated.

Given these tuples, let $\mathbf{D}$ execute A' to obtain a hypothesis h, and sample a fresh set of $p(n)$ samples $((\mathbf{Y}, \mathbf{y}), (\mathbf{z}, z))$ computed as before ($p(n)$ is a polynomial to be defined later). Now, let $\mathbf{D}$ then apply h to all tuples $(\mathbf{Y}, \mathbf{y})$, to obtain a vector $Z^* \in \mathbb{Z}_2^{p(n)(n+1)}$. Let $\mathbf{D}$ output 1 if $\frac{1}{n}|\{j : Z_j^* \neq (\mathbf{z}, z)_j\}| \leq 1/2 - 1/t(n) + 1/2t(n)$ and 0 otherwise.

Now we conduct case analysis for $\mathbf{D}$. Consider the case that $(\mathbf{A}, \mathbf{q})$ given as input to $\mathbf{D}$ is sampled according to the $\text{poly}-\text{DLPN}_{\theta,n}$ distribution. In this case, the dataset set of size $m(n)$ computed by $\mathbf{D}$ is exactly distributed according to $\rho_{\mathcal{Y},\mathcal{Z}}$. Since h satisfies equation (3) with probability at least $1/t(n)$, it follows that in this case $\epsilon \triangleq \frac{1}{n}|\{j : Z_j^* \neq (\mathbf{z}, z)_j\}| \leq 1/2 - 1/2t(n)$ with high probability if we choose $p(n)$ large enough. By application of Chernoff bounds, we can choose $p(n) \triangleq t(n)^3$ such that $\epsilon \leq 1/2 - 1/2t(n)$ with probability at least $1 - \exp(-t(n))$. Hence, in this case $\mathbf{D}$ outputs 1 with probability at least $1/\text{poly}(t(n)) - \exp(-t(n))$ (taking into account $\epsilon \leq 1/2 - 1/2t(n)$ with probability at least $1/t(n)$).

Now, consider the case that $(\mathbf{A}, \mathbf{q})$ given as input to $\mathbf{D}$ is sampled uniformly at random from the domain. In this case, it is clear that, because $\mathbf{q}$ is a uniformly random string, $\epsilon' \triangleq \frac{1}{n}|\{j : Z_j^* \neq (\mathbf{z}, z)_j\}| \geq 1/2 - 1/3t(n)$, with probability at least $1 - \exp(-t(n))$. This follows again by application of Chernoff bounds, since $p(n) \triangleq t(n)^3$. Therefore, the output of $\mathbf{D}$ in this case is 1 with probability at most $\exp(-t(n))$.

This completes the analysis because we have contradicted the equation in definition 3.5.

To end the proof, we consider the easier case of $\rho_{\text{uni}} = \rho_{\mathcal{X},\mathcal{Z}}$. Consider that a sample of $\rho_{\mathcal{X},\mathcal{Z}}$ is of the form $(x = (\mathbf{x}, i), z = (\mathbf{Y}\mathbf{w} + \mathbf{b}', \mathbf{y}\mathbf{w} + \mathbf{b}''))$. By definition of the separation μ—see beginning of section 3.1—the first n bits of z are $\mathbf{Y}\mathbf{w} + \mathbf{b}'$, and can be written as $\mathbf{A}\mathbf{w} + \mathbf{b}'$, for $\mathbf{b}' \sim \text{Ber}(n^{-0.5})^n$, $\mathbf{w} \sim \text{Ber}(n^{-0.5})^n$, and $\mathbf{A}$ uniformly random. Hence, the first n bits of z are a sample from the low-noise LPN distribution, which is actually completely independent of $x = (\mathbf{x}, i)$. Therefore, it is clearly hard to achieve

$$\Pr_{A',\rho\sim\mu}\left[\mathop{\mathbb{E}}_{(x,z)\sim\rho_{\mathcal{X},\mathcal{Z}}}\left[\ell_0(h, x)\right] \leq 1/2 - 1/t(n) : h \leftarrow A'\right] \geq 1/t(n) \tag{4}$$

without refuting the $\text{poly}-\text{LPN}_{\theta,n}$ assumption.

After proving hardness for $\rho_{\text{uni}} \in \{\rho_{\mathcal{X},\mathcal{Z}}, \rho_{\mathcal{Y},\mathcal{Z}}\}$, this suffices to prove the theorem. □

C. Proof of Theorem 4.3

Theorem C.1 (Theorem 4.3 restated)**.** *Suppose that $\mu = (\chi, \eta, \zeta)$ is a super-polynomial computational separation. Then, for any polynomial t, and algorithm D running in time $t(n)$,*

$$\mathbb{E}\Big[D(\text{View}(\text{A} \leftrightarrow \text{B})) = b_B\Big] < 1/2 + 1/t(n)$$

Proof. Towards contradiction, suppose that there exists a probabilistic time $\text{poly}(n)$ algorithm D such that

$$\mathbb{E}\Big[D(\text{View}(\text{A} \leftrightarrow \text{B})) = b_B\Big] \geq 1/2 + 1/\text{poly}(n)$$

We show that this implies that $\mu = (\chi, \eta, \zeta)$ is not a super-polynomial computational separation.

Define the distributions $H_1, \cdots, H_{k+1}$, where H_i is defined as a sample from the following process:

1. Sample $(y_j, z_j) \sim \rho_{\mathcal{Y},\mathcal{Z}}$ for $j \in [k+1]$.
2. For every z_j for $j \in [i, k+1]$, replace it with a random bit σ_j.
3. Output these $k+1$ pairs.

Observe that by definition, when b_B (Bob's bit) is 0, then $\text{View}(\text{A} \leftrightarrow \text{B})$ is distributed identically to H_0 for $\rho \sim \mu$. On the other hand, when b_B is 1, then $\text{View}(\text{A} \leftrightarrow \text{B})$ is distributed identically to H_{k+1} for $\rho \sim \mu$. Thus, the existence of D implies that there exists a probabilistic $\text{poly}(n)$ time decision algorithm D', such that

$$\mathop{\mathbb{E}}_{(y_i,z_i)\sim H_{k+1}}\Big[D'\left((y_i, z_i)_{i\in[k+1]}\right) = 1\Big] - \mathop{\mathbb{E}}_{(y_i,\sigma_i)\sim H_0}\Big[D'\left((y_i, \sigma_i)_{i\in[k+1]}\right) = 1\Big] \geq \frac{1}{\text{poly}(n)}$$

By a standard hybrid argument, this implies that

$$\mathop{\mathbb{E}}_{j\in[k+1]}\Big[\mathop{\mathbb{E}}_{s\sim H_j}\Big[D'(s) = 1\Big] - \mathop{\mathbb{E}}_{s\sim H_{j-1}}\Big[D'(s) = 1\Big]\Big] \geq \frac{1}{(k+1)\text{poly}(n)}$$

The number of examples k can be taken to be $\text{poly}(n)$, by the assumption that Alice's multimodal learning algorithm runs in polynomial time. Thus, we get:

$$\mathop{\mathbb{E}}_{j}\left[\mathop{\mathbb{E}}_{s\sim H_j}\left[D'(s)=1\right]-\mathop{\mathbb{E}}_{s\sim H_{j-1}}\left[D'(s)=1\right]\right]\geq\frac{1}{\text{poly}(n)} \tag{5}$$

Using equation (5), it is possible to derive a randomized prediction algorithm P_μ for μ that satisfies

$$\Pr_{\substack{P_\mu,\rho\sim\mu,\\(x,y,z)\sim\rho}}\left[P_\mu(x,y)=z\right]\geq\frac{1}{2}+\frac{1}{\text{poly}(n)} \tag{6}$$

where the predictor has access to $k\leq\text{poly}(n)$ samples from ρ. Such a randomized predictor is enough to imply that there exists a polynomial $p:\mathbb{N}\to\mathbb{N}$ such that there is a time $p(n)$ probabilistic algorithm A such that, when $\rho\sim\mu$, and given access to $p(n)$ datapoints sampled according to ρ, A outputs a hypothesis that obtains population risk $\ell_{\text{pop}}(h)\leq 1/2-p(n)$ with probability $1/p(n)$ over μ,ρ and randomness of A. This follows from a standard "constructive averaging" argument, which we omit here (see Arora and Barak (2009) (appendix A) and Karchmer (2024) section 5.2 for example). The existence of A as described above completes the proof of the theorem. Hence, let us continue by constructing P_μ that satisfies equation (6).

The main idea is to use the fact that D' can be used to generate evidence that, for a random index j, a label bit b_j is the correct label with respect to the underlying instance of the multimodal learning task. This is because D' should output 0 with slightly higher probability in this case (see equation (5)).

Thus, we define P_μ:

Algorithm 3 $P_\mu \mid \rho\sim\mu$

1: **Input**: (x^*,y^*); $k\leq\text{poly}(n)$ samples from ρ.
2: **Output**: $z^*\in\mathcal{Z}$.
3: Choose uniformly random $j\in[k+1]$.
4: Use k input samples to then sample $s\sim H_j$. Let b_j be the label bit in the j^{th} tuple of s.
5: Derive the set s' from s by replacing x_j,y_j from the j^{th} tuple $(x_j,y_j,b_j)\in s$, with x^*,y^*.
6: **Output** $D'(s')+b_j \mod 2$.

To analyze the probability that P_μ satisfies equation (6), we condition on the correctness of b_j:

$$\Pr\left[P_\mu(x,y)=z\right]=\Pr\left[P_\mu(x,y)=z|b_j=z\right]\cdot\Pr[b_j=z]+\Pr\left[P_\mu(x,y)=z|b_j\neq z\right]\cdot\Pr[b_j\neq z] \tag{7}$$

All the probabilities are taken over $P_\mu,\rho\sim\mu,(x,y,z)\sim\rho$. We then get that, because b_j is by definition a uniformly random bit:

$$\Pr\left[P_\mu(x,y)=z\right]=\frac{1}{2}\left(\Pr\left[P_\mu(x,y)=z|b_j=z\right]+\Pr\left[P_\mu(x,y)=z|b_j\neq z\right]\right) \tag{8}$$

By considering the output of P_μ, we can write this as:

$$\Pr\left[P_\mu(x,y)=z\right]=\frac{1}{2}\left(\Pr\left[D'(s')=0|b_j=z\right]+\Pr\left[D'(s')=1|b_j\neq z\right]\right) \tag{9}$$

$$=\frac{1}{2}\left(1+\Pr\left[D'(s')=0|b_j=z\right]-\Pr\left[D'(s')=0|b_j\neq z\right]\right) \tag{10}$$

$$=\frac{1}{2}+\frac{1}{2}\left(\Pr\left[D'(s')=0|b_j=z\right]-\Pr\left[D'(s')=0|b_j\neq z\right]\right) \tag{11}$$

Then, by knowledge of the bounded difference from equation (5), and by observing that s' (sampled by P_μ) is distributed identically to H_{j-1} (conditioned on $b_j=z$) while s' is distributed identically to H_j (conditioned on $b_j\neq z$), we may conclude that:

$$\Pr\left[P_\mu(x,y)=z\right]\geq\frac{1}{2}+\frac{1}{\text{poly}(n)} \tag{12}$$

as we desired. □