Electronic Colloquium on Computational Complexity, Report No. 121 (2017)

Extractor-Based Time-Space Lower Bounds for Learning

Sumegha Garg* Ran Raz† Avishay Tal‡

Abstract

A matrix $M : A \times X \to \{-1, 1\}$ corresponds to the following learning problem: An unknown element $x \in X$ is chosen uniformly at random. A learner tries to learn x from a stream of samples, $(a_1, b_1), (a_2, b_2) \ldots$, where for every i, $a_i \in A$ is chosen uniformly at random and $b_i = M(a_i, x)$.

Assume that k, ℓ, r are such that any submatrix of M of at least $2^{-k} \cdot |A|$ rows and at least $2^{-\ell} \cdot |X|$ columns, has a bias of at most 2^{-r}. We show that any learning algorithm for the learning problem corresponding to M requires either a memory of size at least $\Omega(k \cdot \ell)$, or at least $2^{\Omega(r)}$ samples. The result holds even if the learner has an exponentially small success probability (of $2^{-\Omega(r)}$).

In particular, this shows that for a large class of learning problems, any learning algorithm requires either a memory of size at least $\Omega((\log |X|) \cdot (\log |A|))$ or an exponential number of samples, achieving a tight $\Omega((\log |X|) \cdot (\log |A|))$ lower bound on the size of the memory, rather than a bound of $\Omega\left(\min\left\{(\log |X|)^2, (\log |A|)^2\right\}\right)$ obtained in previous works [R17, MM17b].

Moreover, our result implies all previous memory-samples lower bounds, as well as a number of new applications.

Our proof builds on [R17] that gave a general technique for proving memory-samples lower bounds.

1 Introduction

Can one prove unconditional lower bounds on the number of samples needed for learning, under memory constraints? The study of the resources needed for learning, under memory constraints was initiated by Shamir [S14] and by Steinhardt, Valiant and Wager [SVW16]. While the main motivation for studying this question comes from learning theory, the problem is also relevant to computational complexity and cryptography [R16, VV16, KRT16].

Steinhardt, Valiant and Wager conjectured that any algorithm for learning parities of size n requires either a memory of size $\Omega(n^2)$ or an exponential number of samples. This conjecture was proven in [R16], showing for the first time a learning problem that is infeasible

*Department of Computer Science, Princeton University.

†Department of Computer Science, Princeton University. Research supported by the Simons Collaboration on Algorithms and Geometry and by the National Science Foundation grant No. CCF-1412958.

‡Institute for Advanced Study, Princeton, NJ. Research supported by the Simons Collaboration on Algorithms and Geometry and by the National Science Foundation grant No. CCF-1412958.

ISSN 1433-8092

under super-linear memory constraints. Building on [R16], it was proved in [KRT16] that learning parities of sparsity ℓ is also infeasible under memory constraints that are super-linear in n, as long as $\ell \geq \omega(\log n / \log\log n)$. Consequently, learning linear-size DNF Formulas, linear-size Decision Trees and logarithmic-size Juntas were all proved to be infeasible under super-linear memory constraints [KRT16] (by a reduction from learning sparse parities).

Can one prove similar memory-samples lower bounds for other learning problems?

As in [R17], we represent a learning problem by a matrix. Let X, A be two finite sets of size larger than 1 (where X represents the concept-class that we are trying to learn and A represents the set of possible samples). Let $M : A \times X \to \{-1, 1\}$ be a matrix. The matrix M represents the following learning problem: An unknown element $x \in X$ was chosen uniformly at random. A learner tries to learn x from a stream of samples, $(a_1, b_1), (a_2, b_2) \ldots$, where for every i, $a_i \in A$ is chosen uniformly at random and $b_i = M(a_i, x)$.

Let $n = \log |X|$ and $n' = \log |A|$.

A general technique for proving memory-samples lower bounds was given in [R17]. The main result of [R17] shows that if the norm of the matrix M is sufficiently small, then any learning algorithm for the corresponding learning problem requires either a memory of size at least $\Omega\left((\min\{n, n'\})^2\right)$, or an exponential number of samples. This gives a general memory-samples lower bound that applies for a large class of learning problems.

Independently of [R17], Moshkovitz and Moshkovitz also gave a general technique for proving memory-samples lower bounds [MM17a]. Their initial result was that if M has a (sufficiently strong) mixing property then any learning algorithm for the corresponding learning problem requires either a memory of size at least $1.25 \cdot \min\{n, n'\}$ or an exponential number of samples [MM17a]. In a recent subsequent work [MM17b], they improved their result, and obtained a theorem that is very similar to the one proved in [R17]. (The result of [MM17b] is stated in terms of a combinatorial mixing property, rather than matrix norm. The two notions are closely related (see in particular Corollary 5.1 and Note 5.1 in [BL06])).

Our Results

The results of [R17] and [MM17b] gave a lower bound of at most $\Omega\left((\min\{n, n'\})^2\right)$ on the size of the memory, whereas the best that one could hope for, in the information theoretic setting (that is, in the setting where the learner's computational power is unbounded), is a lower bound of $\Omega(n \cdot n')$, which may be significantly larger in cases where n is significantly larger than n', or vice versa.

In this work, we build on [R17] and obtain a general memory-samples lower bound that applies for a large class of learning problems and shows that for every problem in that class, any learning algorithm requires either a memory of size at least $\Omega(n \cdot n')$ or an exponential number of samples.

Our result is stated in terms of the properties of the matrix M as a two-source extractor. Two-source extractors, first studied by Santha and Vazirani [SV84] and Chor and Goldreich [CG88], are central objects in the study of randomness and derandomization. We show that even a relatively weak two-source extractor implies a relatively strong memory-samples lower bound. We note that two-source extractors have been extensively studied in numerous of works and there are known techniques for proving that certain matrices are relatively good two-source extractors.

Our main result can be stated as follows (Corollary 3): Assume that k, ℓ, r are such that any submatrix of M of at least $2^{-k} \cdot |A|$ rows and at least $2^{-\ell} \cdot |X|$ columns, has a bias of at most 2^{-r}. Then, any learning algorithm for the learning problem corresponding to M requires either a memory of size at least $\Omega(k \cdot \ell)$, or at least $2^{\Omega(r)}$ samples. The result holds even if the learner has an exponentially small success probability (of $2^{-\Omega(r)}$).

A more detailed result, in terms of the constants involved, is stated in Theorem 1 in terms of the properties of M as an L_2-Extractor, a new notion that we define in Definition 2.1, and is closely related to the notion of two-source extractor. (The two notions are equivalent up to small changes in the parameters.)

All of our results (and all applications) hold even if the learner is only required to *weakly learn* x, that is to output a hypothesis $h : A \to \{-1, 1\}$ with a non-negligible correlation with the x-th column of the matrix M. We prove in Theorem 2 that even if the learner is only required to output a hypothesis that agrees with the x-th column of M on more than a $1/2 + 2^{-\Omega(r)}$ fraction of the rows, the success probability is at most $2^{-\Omega(r)}$.

As in [R16, KRT16, R17], we model the learning algorithm by a *branching program*. A branching program is the strongest and most general model to use in this context. Roughly speaking, the model allows a learner with infinite computational power, and bounds only the memory size of the learner and the number of samples used.

As mentioned above, our result implies all previous memory-samples lower bounds, as well as new applications. In particular:

1. **Parities:** A learner tries to learn $x = (x_1, \ldots, x_n) \in \{0,1\}^n$, from random linear equations over $\mathbb{F}_2$. It was proved in [R16] (and follows also from [R17]) that any learning algorithm requires either a memory of size $\Omega(n^2)$ or an exponential number of samples. The same result follows by Corollary 3 and the fact that inner product is a good two-source extractor [CG88].

2. **Sparse parities:** A learner tries to learn $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ of sparsity ℓ, from random linear equations over $\mathbb{F}_2$. In Section 5.2, we reprove the main results of [KRT16]. In particular, any learning algorithm requires:

   (a) Assuming $\ell \leq n/2$: either a memory of size $\Omega(n \cdot \ell)$ or $2^{\Omega(\ell)}$ samples.

   (b) Assuming $\ell \leq n^{0.9}$: either a memory of size $\Omega(n \cdot \ell^{0.99})$ or $\ell^{\Omega(\ell)}$ samples.

3. **Learning from sparse linear equations:** A learner tries to learn $x = (x_1, \ldots, x_n) \in \{0,1\}^n$, from random sparse linear equations, of sparsity ℓ, over $\mathbb{F}_2$. In Section 5.3, we prove that any learning algorithm requires:

   (a) Assuming $\ell \leq n/2$: either a memory of size $\Omega(n \cdot \ell)$ or $2^{\Omega(\ell)}$ samples.

   (b) Assuming $\ell \leq n^{0.9}$: either a memory of size $\Omega(n \cdot \ell^{0.99})$ or $\ell^{\Omega(\ell)}$ samples.

4. **Learning from low-degree equations:** A learner tries to learn $x = (x_1, \ldots, x_n) \in \{0,1\}^n$, from random multilinear polynomial equations of degree at most d, over $\mathbb{F}_2$. In Section 5.4, we prove that if $d \leq 0.99 \cdot n$, any learning algorithm requires either a memory of size $\Omega\left(\binom{n}{\leq d} \cdot n/d\right)$ or $2^{\Omega(n/d)}$ samples.

5. **Low-degree polynomials:** A learner tries to learn an n'-variate multilinear polynomial p of degree at most d over $\mathbb{F}_2$, from random evaluations of p over $\mathbb{F}_2^{n'}$. In Section 5.5, we prove that if $d \leq 0.99 \cdot n'$, any learning algorithm requires either a memory of size $\Omega\left(\binom{n'}{\leq d} \cdot n'/d\right)$ or $2^{\Omega(n'/d)}$ samples.

6. **Error-correcting codes:** A learner tries to learn a codeword from random coordinates: Assume that $M : A \times X \to \{-1, 1\}$ is such that for some $|X|^{-1} \leq \epsilon < 1$, any pair of different columns of M, agree on at least $\frac{1-\epsilon}{2} \cdot |A|$ and at most $\frac{1+\epsilon}{2} \cdot |A|$ coordinates. In Section 5.6, we prove that any learning algorithm for the learning problem corresponding to M requires either a memory of size $\Omega\big((\log|X|) \cdot (\log(1/\epsilon))\big)$ or $\left(\frac{1}{\epsilon}\right)^{\Omega(1)}$ samples. We also point to a relation between our results and statistical-query dimension [K98, BFJKMR94].

7. **Random matrices:** Let X, A be finite sets, such that, $|A| \geq (2\log|X|)^{10}$ and $|X| \geq (2\log|A|)^{10}$. Let $M : A \times X \to \{-1, 1\}$ be a random matrix. Fix $k = \frac{1}{2}\log|A|$ and $\ell = \frac{1}{2}\log|X|$. With very high probability, any submatrix of M of at least $2^{-k} \cdot |A|$ rows and at least $2^{-\ell} \cdot |X|$ columns, has a bias of at most $2^{-\Omega(\min\{k,\ell\})}$. Thus, by Corollary 3, any learning algorithm for the learning problem corresponding to M requires either a memory of size $\Omega\left((\log|X|) \cdot (\log|A|)\right)$, or $\left(\min\{|X|, |A|\}\right)^{\Omega(1)}$ samples.

We note also that our results about learning from sparse linear equations have applications in bounded-storage cryptography. This is similar to [R16, KRT16], but in a different range of the parameters. In particular, for every $\omega(\log n) \leq \ell \leq n$, our results give an encryption scheme that requires a private key of length n, and time complexity of $O(\ell \log n)$ per encryption/decryption of each bit, using a random access machine. The scheme is provenly and unconditionally secure as long as the attacker uses at most $o(n\ell)$ memory bits and the scheme is used at most $2^{o(\ell)}$ times.

Techniques

Our proof follows the lines of the proof of [R17] and builds on that proof. The proof of [R17] considered the norm of the matrix M, and thus essentially reduced the entire matrix to only one parameter. In our proof, we consider the properties of M as a two-source extractor, and hence we have three parameters (k, ℓ, r), rather than one. Considering these three parameters, rather than one, enables a more refined analysis, resulting in a stronger lower bound with a slightly simpler proof.

A proof outline is given in Section 3.

Motivation and Discussion

Many previous works studied the resources needed for learning, under certain information, communication or memory constraints (see in particular [S14, SVW16, R16, VV16, KRT16, MM17a, R17, MT17, MM17b] and the many references given there). A main message of some of these works is that for some learning problems, access to a relatively large memory is crucial. In other words, in some cases, learning is infeasible, due to memory constraints.

From the point of view of human learning, such results may help to explain the importance of memory in cognitive processes. From the point of view of machine learning, these results imply that a large class of learning algorithms cannot learn certain concept classes. In particular, this applies to any bounded-memory learning algorithm that considers the samples one by one. In addition, these works are related to computational complexity and have applications in cryptography.

Related Work

Independently of our work, Beame, Oveis Gharan and Yang also gave a combinatorial property of a matrix M, that holds for a large class of matrices and implies that any learning algorithm for the corresponding learning problem requires either a memory of size $\Omega\left(\left(\log|X|\right)\cdot\left(\log|A|\right)\right)$ or an exponential number of samples (when $|A| \leq |X|$) [BOGY17]. Their property is based on a measure of how matrices amplify the 2-norms of probability distributions that is more refined than the 2-norms of these matrices. Their proof also builds on [R17].

They also show, as an application, tight time-space lower bounds for learning low-degree polynomials, as well as other applications.

2 Preliminaries

Denote by $\mathcal{U}_X : X \to \mathbb{R}^+$ the uniform distribution over X. Denote by log the logarithm to base 2. Denote by $\binom{n}{\leq k} = \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{k}$.

For a random variable Z and an event E, we denote by $\mathbb{P}_Z$ the distribution of the random variables Z, and we denote by $\mathbb{P}_{Z|E}$ the distribution of the random variable Z conditioned on the event E.

Viewing a Learning Problem as a Matrix

Let X, A be two finite sets of size larger than 1. Let $n = \log_2|X|$.

Let $M : A \times X \to \{-1, 1\}$ be a matrix. The matrix M corresponds to the following learning problem: There is an unknown element $x \in X$ that was chosen uniformly at random. A learner tries to learn x from samples (a, b), where $a \in A$ is chosen uniformly at random and $b = M(a, x)$. That is, the learning algorithm is given a stream of samples, $(a_1, b_1), (a_2, b_2) \ldots$, where each a_t is uniformly distributed and for every t, $b_t = M(a_t, x)$.

Norms and Inner Products

Let $p \geq 1$. For a function $f : X \to \mathbb{R}$, denote by $\|f\|_p$ the ℓ_p norm of f, with respect to the uniform distribution over X, that is:

$$\|f\|_p = \left(\mathop{\mathbf{E}}_{x \in_R X}\left[|f(x)|^p\right]\right)^{1/p}.$$

For two functions $f, g : X \to \mathbb{R}$, define their inner product with respect to the uniform distribution over X as

$$\langle f, g\rangle = \mathop{\mathbf{E}}_{x \in_R X}[f(x) \cdot g(x)].$$

For a matrix $M : A \times X \to \mathbb{R}$ and a row $a \in A$, we denote by $M_a : X \to \mathbb{R}$ the function corresponding to the a-th row of M. Note that for a function $f : X \to \mathbb{R}$, we have $\langle M_a, f \rangle = \frac{(M \cdot f)_a}{|X|}$.

L_2-Extractors and L_∞-Extractors

Definition 2.1. **L_2-Extractor:** *Let X, A be two finite sets. A matrix $M : A \times X \to \{-1, 1\}$ is a (k, ℓ)-L_2-Extractor with error 2^{-r}, if for every non-negative $f : X \to \mathbb{R}$ with $\frac{\|f\|_2}{\|f\|_1} \leq 2^\ell$ there are at most $2^{-k} \cdot |A|$ rows a in A with*

$$\frac{|\langle M_a, f \rangle|}{\|f\|_1} \geq 2^{-r} .$$

Let Ω be a finite set. We denote a distribution over Ω as a function $f : \Omega \to \mathbb{R}^+$ such that $\sum_{x \in \Omega} f(x) = 1$. We say that a distribution $f : \Omega \to \mathbb{R}^+$ has min-entropy k if for all $x \in \Omega$, we have $f(x) \leq 2^{-k}$.

Definition 2.2. **$L_\infty-$Extractor:** *Let X, A be two finite sets. A matrix $M : A \times X \to \{-1, 1\}$ is a $(k, \ell \sim r)$-L_∞-Extractor if for every distribution $p_x : X \to \mathbb{R}^+$ with min-entropy at least $(\log(|X|) - \ell)$ and every distribution $p_a : A \to \mathbb{R}^+$ with min-entropy at least $(\log(|A|) - k)$,*

$$\left| \sum_{a' \in A} \sum_{x' \in X} p_a(a') \cdot p_x(x') \cdot M(a', x') \right| \leq 2^{-r}.$$

Branching Program for a Learning Problem

In the following definition, we model the learner for the learning problem that corresponds to the matrix M, by a *branching program*.

Definition 2.3. Branching Program for a Learning Problem: *A branching program of length m and width d, for learning, is a directed (multi) graph with vertices arranged in $m+1$ layers containing at most d vertices each. In the first layer, that we think of as layer 0, there is only one vertex, called the start vertex. A vertex of outdegree 0 is called a leaf. All vertices in the last layer are leaves (but there may be additional leaves). Every non-leaf vertex in the program has $2|A|$ outgoing edges, labeled by elements $(a, b) \in A \times \{-1, 1\}$, with exactly one edge labeled by each such (a, b), and all these edges going into vertices in the next layer. Each leaf v in the program is labeled by an element $\tilde{x}(v) \in X$, that we think of as the output of the program on that leaf.*

Computation-Path: *The samples $(a_1, b_1), \ldots, (a_m, b_m) \in A \times \{-1, 1\}$ that are given as input, define a computation-path in the branching program, by starting from the start vertex and following at step t the edge labeled by (a_t, b_t), until reaching a leaf. The program outputs the label $\tilde{x}(v)$ of the leaf v reached by the computation-path.*

Success Probability: *The success probability of the program is the probability that $\tilde{x} = x$, where $\tilde{x}$ is the element that the program outputs, and the probability is over $x, a_1, \ldots, a_m$ (where x is uniformly distributed over X and $a_1, \ldots, a_m$ are uniformly distributed over A, and for every t, $b_t = M(a_t, x)$).*

3 Overview of the Proof

The proof follows the lines of the proof of [R17] and builds on that proof.

Assume that M is a (k,ℓ)-L_2-extractor with error $2^{-r'}$, and let $r = \min\{k, \ell, r'\}$. Let B be a branching program for the learning problem that corresponds to the matrix M. Assume for a contradiction that B is of length $m = 2^{\epsilon r}$ and width $d = 2^{\epsilon k\ell}$, where ϵ is a small constant.

We define the *truncated-path*, $\mathcal{T}$, to be the same as the computation-path of B, except that it sometimes stops before reaching a leaf. Roughly speaking, $\mathcal{T}$ stops before reaching a leaf if certain "bad" events occur. Nevertheless, we show that the probability that $\mathcal{T}$ stops before reaching a leaf is negligible, so we can think of $\mathcal{T}$ as almost identical to the computation-path.

For a vertex v of B, we denote by E_v the event that $\mathcal{T}$ reaches the vertex v. We denote by $\Pr(v) = \Pr(E_v)$ the probability for E_v (where the probability is over $x, a_1, \ldots, a_m$), and we denote by $\mathbb{P}_{x|v} = \mathbb{P}_{x|E_v}$ the distribution of the random variable x conditioned on the event E_v. Similarly, for an edge e of the branching program B, let E_e be the event that $\mathcal{T}$ traverses the edge e. Denote, $\Pr(e) = \Pr(E_e)$, and $\mathbb{P}_{x|e} = \mathbb{P}_{x|E_e}$.

A vertex v of B is called *significant* if

$$\left\|\mathbb{P}_{x|v}\right\|_2 > 2^{\ell} \cdot 2^{-n}.$$

Roughly speaking, this means that conditioning on the event that $\mathcal{T}$ reaches the vertex v, a non-negligible amount of information is known about x. In order to guess x with a non-negligible success probability, $\mathcal{T}$ must reach a significant vertex. Lemma 4.1 shows that the probability that $\mathcal{T}$ reaches any significant vertex is negligible, and thus the main result follows.

To prove Lemma 4.1, we show that for every fixed significant vertex s, the probability that $\mathcal{T}$ reaches s is at most $2^{-\Omega(k\ell)}$ (which is smaller than one over the number of vertices in B). Hence, we can use a union bound to prove the lemma.

The proof that the probability that $\mathcal{T}$ reaches s is extremely small is the main part of the proof. To that end, we use the following functions to measure the progress made by the branching program towards reaching s.

Let L_i be the set of vertices v in layer-i of B, such that $\Pr(v) > 0$. Let Γ_i be the set of edges e from layer-$(i-1)$ of B to layer-i of B, such that $\Pr(e) > 0$. Let

$$\mathcal{Z}_i = \sum_{v \in L_i} \Pr(v) \cdot \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^k,$$

$$\mathcal{Z}'_i = \sum_{e \in \Gamma_i} \Pr(e) \cdot \langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^k.$$

We think of $\mathcal{Z}_i, \mathcal{Z}'_i$ as measuring the progress made by the branching program, towards reaching a state with distribution similar to $\mathbb{P}_{x|s}$.

We show that each $\mathcal{Z}_i$ may only be negligibly larger than $\mathcal{Z}_{i-1}$. Hence, since it's easy to calculate that $\mathcal{Z}_0 = 2^{-2nk}$, it follows that $\mathcal{Z}_i$ is close to 2^{-2nk}, for every i. On the other hand, if s is in layer-i then $\mathcal{Z}_i$ is at least $\Pr(s) \cdot \langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle^k$. Thus, $\Pr(s) \cdot \langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle^k$ cannot be much larger than 2^{-2nk}. Since s is significant, $\langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle^k > 2^{\ell k} \cdot 2^{-2nk}$ and hence $\Pr(s)$ is at most $2^{-\Omega(k\ell)}$.

The proof that $\mathcal{Z}_i$ may only be negligibly larger than $\mathcal{Z}_{i-1}$ is done in two steps: Claim 4.12 shows by a simple convexity argument that $\mathcal{Z}_i \leq \mathcal{Z}'_i$. The hard part, that is done in Claim 4.10 and Claim 4.11, is to prove that $\mathcal{Z}'_i$ may only be negligibly larger than $\mathcal{Z}_{i-1}$.

For this proof, we define for every vertex v, the set of edges $\Gamma_{out}(v)$ that are going out of v, such that $\Pr(e) > 0$. Claim 4.10 shows that for every vertex v,

$$\sum_{e \in \Gamma_{out}(v)} \Pr(e) \cdot \langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^k$$

may only be negligibly higher than

$$\Pr(v) \cdot \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^k.$$

For the proof of Claim 4.10, which is the hardest proof in the paper, and the most important place where our proof deviates from (and simplifies) the proof of [R17], we consider the function $\mathbb{P}_{x|v} \cdot \mathbb{P}_{x|s}$. We first show how to bound $\left\|\mathbb{P}_{x|v} \cdot \mathbb{P}_{x|s}\right\|_2$. We then consider two cases: If $\left\|\mathbb{P}_{x|v} \cdot \mathbb{P}_{x|s}\right\|_1$ is negligible, then $\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^k$ is negligible and doesn't contribute much, and we show that for every $e \in \Gamma_{out}(v)$, $\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^k$ is also negligible and doesn't contribute much. If $\left\|\mathbb{P}_{x|v} \cdot \mathbb{P}_{x|s}\right\|_1$ is non-negligible, we use the bound on $\left\|\mathbb{P}_{x|v} \cdot \mathbb{P}_{x|s}\right\|_2$ and the assumption that M is a (k, ℓ)-L_2-extractor to show that for almost all edges $e \in \Gamma_{out}(v)$, we have that $\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^k$ is very close to $\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^k$. Only an exponentially small (2^{-k}) fraction of edges are "bad" and give a significantly larger $\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^k$.

The reason that in the definitions of $\mathcal{Z}_i$ and $\mathcal{Z}'_i$ we raised $\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle$ and $\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle$ to the power of k is that this is the largest power for which the contribution of the "bad" edges is still small (as their fraction is 2^{-k}).

This outline oversimplifies many details. Let us briefly mention two of them. First, it is not so easy to bound $\left\|\mathbb{P}_{x|v} \cdot \mathbb{P}_{x|s}\right\|_2$. We do that by bounding $\left\|\mathbb{P}_{x|s}\right\|_2$ and $\left\|\mathbb{P}_{x|v}\right\|_\infty$. In order to bound $\left\|\mathbb{P}_{x|s}\right\|_2$, we force $\mathcal{T}$ to stop whenever it reaches a significant vertex (and thus we are able to bound $\left\|\mathbb{P}_{x|v}\right\|_2$ for every vertex reached by $\mathcal{T}$). In order to bound $\left\|\mathbb{P}_{x|v}\right\|_\infty$, we force $\mathcal{T}$ to stop whenever $\mathbb{P}_{x|v}(x)$ is large, which allows us to consider only the "bounded" part of $\mathbb{P}_{x|v}$. (This is related to the technique of *flattening* a distribution that was used in [KR13]). Second, some edges are so "bad" that their contribution to $\mathcal{Z}'_i$ is huge so they cannot be ignored. We force $\mathcal{T}$ to stop before traversing any such edge. (This is related to an idea that was used in [KRT16] of analyzing separately paths that traverse "bad" edges). We show that the total probability that $\mathcal{T}$ stops before reaching a leaf is negligible.

4 Main Result

Theorem 1. *Let $\frac{1}{100} < c < \frac{2}{3}$. Fix γ to be such that $\frac{3c}{2} < \gamma^2 < 1$.*

Let X, A be two finite sets. Let $n = \log_2 |X|$. Let $M : A \times X \to \{-1, 1\}$ be a matrix which is a (k', ℓ')-L_2-extractor with error $2^{-r'}$, for sufficiently large[1] k', ℓ' and r', where $\ell' \leq n$. Let

$$r := \min\left\{\frac{r'}{2}, \frac{(1-\gamma)k'}{2}, \frac{(1-\gamma)\ell'}{2} - 1\right\}. \tag{1}$$

[1] By *"sufficiently large"* we mean that k', ℓ', r' are larger than some constant that depends on γ.

Let B be a branching program of length at most 2^r and width at most $2^{c\cdot k'\cdot \ell'}$ for the learning problem that corresponds to the matrix M. Then, the success probability of B is at most $O(2^{-r})$.

Proof. Let

$$k := \gamma k' \qquad \text{and} \qquad \ell := \gamma \ell'/3. \tag{2}$$

Note that by the assumption that k', ℓ' and r' are sufficiently large, we get that k, ℓ and r are also sufficiently large. Since $\ell' \leq n$, we have $\ell + r \leq \frac{\gamma\ell'}{3} + \frac{(1-\gamma)\ell'}{2} < \frac{\ell'}{2} \leq \frac{n}{2}$. Thus,

$$r < n/2 - \ell. \tag{3}$$

Let B be a branching program of length $m = 2^r$ and width $d = 2^{c\cdot k'\cdot \ell'}$ for the learning problem that corresponds to the matrix M. We will show that the success probability of B is at most $O(2^{-r})$.

4.1 The Truncated-Path and Additional Definitions and Notation

We will define the **truncated-path**, $\mathcal{T}$, to be the same as the computation-path of B, except that it sometimes stops before reaching a leaf. Formally, we define $\mathcal{T}$, together with several other definitions and notations, by induction on the layers of the branching program B.

Assume that we already defined the truncated-path $\mathcal{T}$, until it reaches layer-i of B. For a vertex v in layer-i of B, let E_v be the event that $\mathcal{T}$ reaches the vertex v. For simplicity, we denote by $\Pr(v) = \Pr(E_v)$ the probability for E_v (where the probability is over $x, a_1, \ldots, a_m$), and we denote by $\mathbb{P}_{x|v} = \mathbb{P}_{x|E_v}$ the distribution of the random variable x conditioned on the event E_v.

There will be three cases in which the truncated-path $\mathcal{T}$ stops on a non-leaf v:

1. If v is a, so called, significant vertex, where the ℓ_2 norm of $\mathbb{P}_{x|v}$ is non-negligible. (Intuitively, this means that conditioned on the event that $\mathcal{T}$ reaches v, a non-negligible amount of information is known about x).

2. If $\mathbb{P}_{x|v}(x)$ is non-negligible. (Intuitively, this means that conditioned on the event that $\mathcal{T}$ reaches v, the correct element x could have been guessed with a non-negligible probability).

3. If $(M \cdot \mathbb{P}_{x|v})(a_{i+1})$ is non-negligible. (Intuitively, this means that $\mathcal{T}$ is about to traverse a "bad" edge, which is traversed with a non-negligibly higher or lower probability than other edges).

Next, we describe these three cases more formally.

Significant Vertices

We say that a vertex v in layer-i of B is **significant** if

$$\left\|\mathbb{P}_{x|v}\right\|_2 > 2^{\ell} \cdot 2^{-n}.$$

Significant Values

Even if v is not significant, $\mathbb{P}_{x|v}$ may have relatively large values. For a vertex v in layer-i of B, denote by $\mathrm{Sig}(v)$ the set of all $x' \in X$, such that,

$$\mathbb{P}_{x|v}(x') > 2^{2\ell+2r} \cdot 2^{-n}.$$

Bad Edges

For a vertex v in layer-i of B, denote by $\mathrm{Bad}(v)$ the set of all $\alpha \in A$, such that,

$$\left|(M \cdot \mathbb{P}_{x|v})(\alpha)\right| \geq 2^{-r'}.$$

The Truncated-Path $\mathcal{T}$

We define $\mathcal{T}$ by induction on the layers of the branching program B. Assume that we already defined $\mathcal{T}$ until it reaches a vertex v in layer-i of B. The path $\mathcal{T}$ stops on v if (at least) one of the following occurs:

1. v is significant.
2. $x \in \mathrm{Sig}(v)$.
3. $a_{i+1} \in \mathrm{Bad}(v)$.
4. v is a leaf.

Otherwise, $\mathcal{T}$ proceeds by following the edge labeled by (a_{i+1}, b_{i+1}) (same as the computational-path).

4.2 Proof of Theorem 1

Since $\mathcal{T}$ follows the computation-path of B, except that it sometimes stops before reaching a leaf, the success probability of B is bounded (from above) by the probability that $\mathcal{T}$ stops before reaching a leaf, plus the probability that $\mathcal{T}$ reaches a leaf v and $\tilde{x}(v) = x$.

The main lemma needed for the proof of Theorem 1 is Lemma 4.1 that shows that the probability that $\mathcal{T}$ reaches a significant vertex is at most $O(2^{-r})$.

Lemma 4.1. *The probability that $\mathcal{T}$ reaches a significant vertex is at most $O(2^{-r})$.*

Lemma 4.1 is proved in Section 4.3. We will now show how the proof of Theorem 1 follows from that lemma.

Lemma 4.1 shows that the probability that $\mathcal{T}$ stops on a non-leaf vertex, because of the first reason (i.e., that the vertex is significant), is small. The next two lemmas imply that the probabilities that $\mathcal{T}$ stops on a non-leaf vertex, because of the second and third reasons, are also small.

Claim 4.2. *If v is a non-significant vertex of B then*

$$\Pr_x[x \in \mathrm{Sig}(v) \mid E_v] \leq 2^{-2r}.$$

Proof. Since v is not significant,

$$\mathop{\mathbf{E}}_{x' \sim \mathbb{P}_{x|v}} \left[\mathbb{P}_{x|v}(x') \right] = \sum_{x' \in X} \left[\mathbb{P}_{x|v}(x')^2 \right] = 2^n \cdot \mathop{\mathbf{E}}_{x' \in_R X} \left[\mathbb{P}_{x|v}(x')^2 \right] \leq 2^{2\ell} \cdot 2^{-n}.$$

Hence, by Markov's inequality,

$$\Pr_{x' \sim \mathbb{P}_{x|v}} \left[\mathbb{P}_{x|v}(x') > 2^{2r} \cdot 2^{2\ell} \cdot 2^{-n} \right] \leq 2^{-2r}.$$

Since conditioned on E_v, the distribution of x is $\mathbb{P}_{x|v}$, we obtain

$$\Pr_x \left[x \in \mathrm{Sig}(v) \;\middle|\; E_v \right] = \Pr_x \left[\left(\mathbb{P}_{x|v}(x) > 2^{2r} \cdot 2^{2\ell} \cdot 2^{-n} \right) \;\middle|\; E_v \right] \leq 2^{-2r}. \qquad \square$$

Claim 4.3. *If v is a non-significant vertex of B then*

$$\Pr_{a_{i+1}} \left[a_{i+1} \in \mathrm{Bad}(v) \right] \leq 2^{-2r}.$$

Proof. Since v is not significant, $\left\| \mathbb{P}_{x|v} \right\|_2 \leq 2^{\ell} \cdot 2^{-n}$. Since $\mathbb{P}_{x|v}$ is a distribution, $\left\| \mathbb{P}_{x|v} \right\|_1 = 2^{-n}$. Thus,

$$\frac{\left\| \mathbb{P}_{x|v} \right\|_2}{\left\| \mathbb{P}_{x|v} \right\|_1} \leq 2^{\ell} \leq 2^{\ell'}.$$

Since M is a (k', ℓ')-L_2-extractor with error $2^{-r'}$, there are at most $2^{-k'} \cdot |A|$ elements $\alpha \in A$ with

$$\left| \langle M_\alpha, \mathbb{P}_{x|v} \rangle \right| \geq 2^{-r'} \cdot \left\| \mathbb{P}_{x|v} \right\|_1 = 2^{-r'} \cdot 2^{-n}$$

The claim follows since a_{i+1} is uniformly distributed over A and since $k' \geq 2r$ (Equation (1)). $\square$

We can now use Lemma 4.1, Claim 4.2 and Claim 4.3 to prove that the probability that $\mathcal{T}$ stops before reaching a leaf is at most $O(2^{-r})$. Lemma 4.1 shows that the probability that $\mathcal{T}$ reaches a significant vertex and hence stops because of the first reason, is at most $O(2^{-r})$. Assuming that $\mathcal{T}$ doesn't reach any significant vertex (in which case it would have stopped because of the first reason), Claim 4.2 shows that in each step, the probability that $\mathcal{T}$ stops because of the second reason, is at most 2^{-2r}. Taking a union bound over the $m = 2^r$ steps, the total probability that $\mathcal{T}$ stops because of the second reason, is at most 2^{-r}. In the same way, assuming that $\mathcal{T}$ doesn't reach any significant vertex (in which case it would have stopped because of the first reason), Claim 4.3 shows that in each step, the probability that $\mathcal{T}$ stops because of the third reason, is at most 2^{-2r}. Again, taking a union bound over the 2^r steps, the total probability that $\mathcal{T}$ stops because of the third reason, is at most 2^{-r}. Thus, the total probability that $\mathcal{T}$ stops (for any reason) before reaching a leaf is at most $O(2^{-r})$.

Recall that if $\mathcal{T}$ doesn't stop before reaching a leaf, it just follows the computation-path of B. Recall also that by Lemma 4.1, the probability that $\mathcal{T}$ reaches a significant leaf is at most $O(2^{-r})$. Thus, to bound (from above) the success probability of B by $O(2^{-r})$, it remains to bound the probability that $\mathcal{T}$ reaches a non-significant leaf v and $\tilde{x}(v) = x$. Claim 4.4 shows that for any non-significant leaf v, conditioned on the event that $\mathcal{T}$ reaches v, the probability for $\tilde{x}(v) = x$ is at most 2^{-r}, which completes the proof of Theorem 1.

Claim 4.4. *If v is a non-significant leaf of B then*

$$\Pr[\tilde{x}(v) = x \mid E_v] \leq 2^{-r}.$$

Proof. Since v is not significant,

$$\mathop{\mathbf{E}}_{x' \in_R X} \left[\mathbb{P}_{x|v}(x')^2\right] \leq 2^{2\ell} \cdot 2^{-2n}.$$

Hence, for every $x' \in X$,

$$\Pr[x = x' \mid E_v] = \mathbb{P}_{x|v}(x') \leq 2^{\ell} \cdot 2^{-n/2} \leq 2^{-r}$$

since $r \leq n/2 - \ell$ (Equation (3)). In particular,

$$\Pr[\tilde{x}(v) = x \mid E_v] \leq 2^{-r}. \qquad \square$$

This completes the proof of Theorem 1. $\square$

4.3 Proof of Lemma 4.1

Proof. We need to prove that the probability that $\mathcal{T}$ reaches any significant vertex is at most $O(2^{-r})$. Let s be a significant vertex of B. We will bound from above the probability that $\mathcal{T}$ reaches s, and then use a union bound over all significant vertices of B. Interestingly, the upper bound on the width of B is used only in the union bound.

The Distributions $\mathbb{P}_{x|v}$ and $\mathbb{P}_{x|e}$

Recall that for a vertex v of B, we denote by E_v the event that $\mathcal{T}$ reaches the vertex v. For simplicity, we denote by $\Pr(v) = \Pr(E_v)$ the probability for E_v (where the probability is over $x, a_1, \ldots, a_m$), and we denote by $\mathbb{P}_{x|v} = \mathbb{P}_{x|E_v}$ the distribution of the random variable x conditioned on the event E_v.

Similarly, for an edge e of the branching program B, let E_e be the event that $\mathcal{T}$ traverses the edge e. Denote, $\Pr(e) = \Pr(E_e)$ (where the probability is over $x, a_1, \ldots, a_m$), and $\mathbb{P}_{x|e} = \mathbb{P}_{x|E_e}$.

Claim 4.5. *For any edge $e = (v, u)$ of B, labeled by (a, b), such that $\Pr(e) > 0$, for any $x' \in X$,*

$$\mathbb{P}_{x|e}(x') = \begin{cases} 0 & \text{if} \quad x' \in \mathrm{Sig}(v) \quad \text{or} \quad M(a, x') \neq b \\ \mathbb{P}_{x|v}(x') \cdot c_e^{-1} & \text{if} \quad x' \notin \mathrm{Sig}(v) \quad \text{and} \quad M(a, x') = b \end{cases}$$

where c_e is a normalization factor that satisfies,

$$c_e \geq \tfrac{1}{2} - 2 \cdot 2^{-2r}.$$

Proof. Let $e = (v, u)$ be an edge of B, labeled by (a, b), and such that $\Pr(e) > 0$. Since $\Pr(e) > 0$, the vertex v is not significant (as otherwise $\mathcal{T}$ always stops on v and hence $\Pr(e) = 0$). Also, since $\Pr(e) > 0$, we know that $a \notin \mathrm{Bad}(v)$ (as otherwise $\mathcal{T}$ never traverses e and hence $\Pr(e) = 0$).

If $\mathcal{T}$ reaches v, it traverses the edge e if and only if: $x \notin \mathrm{Sig}(v)$ (as otherwise $\mathcal{T}$ stops on v) and $M(a, x) = b$ and $a_{i+1} = a$. Therefore, for any $x' \in X$,

$$\mathbb{P}_{x|e}(x') = \begin{cases} 0 & \text{if} \quad x' \in \mathrm{Sig}(v) \quad \text{or} \quad M(a, x') \neq b \\ \mathbb{P}_{x|v}(x') \cdot c_e^{-1} & \text{if} \quad x' \notin \mathrm{Sig}(v) \quad \text{and} \quad M(a, x') = b \end{cases}$$

where c_e is a normalization factor, given by

$$c_e = \sum_{\{x' \,:\, x' \notin \mathrm{Sig}(v) \,\wedge\, M(a,x')=b\}} \mathbb{P}_{x|v}(x') \;=\; \Pr_x[(x \notin \mathrm{Sig}(v)) \wedge (M(a,x)=b) \mid E_v].$$

Since v is not significant, by Claim 4.2,

$$\Pr_x[x \in \mathrm{Sig}(v) \mid E_v] \leq 2^{-2r}.$$

Since $a \notin \mathrm{Bad}(v)$,

$$\left|\Pr_x[M(a,x)=1 \mid E_v] - \Pr_x[M(a,x)=-1 \mid E_v]\right| = \left|(M \cdot \mathbb{P}_{x|v})(a)\right| \leq 2^{-r'},$$

and hence

$$\Pr_x[M(a,x) \neq b \mid E_v] \leq \tfrac{1}{2} + 2^{-r'}.$$

Hence, by the union bound,

$$c_e = \Pr_x[(x \notin \mathrm{Sig}(v)) \wedge (M(a,x)=b) \mid E_v] \geq \tfrac{1}{2} - 2^{-r'} - 2^{-2r} \geq \tfrac{1}{2} - 2 \cdot 2^{-2r}$$

(where the last inequality follows since $r \leq r'/2$, by Equation (1)). □

Bounding the Norm of $\mathbb{P}_{x|s}$

We will show that $\left\|\mathbb{P}_{x|s}\right\|_2$ cannot be too large. Towards this, we will first prove that for every edge e of B that is traversed by $\mathcal{T}$ with probability larger than zero, $\left\|\mathbb{P}_{x|e}\right\|_2$ cannot be too large.

Claim 4.6. *For any edge e of B, such that $\Pr(e) > 0$,*

$$\left\|\mathbb{P}_{x|e}\right\|_2 \leq 4 \cdot 2^{\ell} \cdot 2^{-n}.$$

Proof. Let $e = (v,u)$ be an edge of B, labeled by (a,b), and such that $\Pr(e) > 0$. Since $\Pr(e) > 0$, the vertex v is not significant (as otherwise $\mathcal{T}$ always stops on v and hence $\Pr(e) = 0$). Thus,

$$\left\|\mathbb{P}_{x|v}\right\|_2 \leq 2^{\ell} \cdot 2^{-n}.$$

By Claim 4.5, for any $x' \in X$,

$$\mathbb{P}_{x|e}(x') = \begin{cases} 0 & \text{if} \quad x' \in \mathrm{Sig}(v) \quad \text{or} \quad M(a,x') \neq b \\ \mathbb{P}_{x|v}(x') \cdot c_e^{-1} & \text{if} \quad x' \notin \mathrm{Sig}(v) \quad \text{and} \quad M(a,x') = b \end{cases}$$

where c_e satisfies,

$$c_e \geq \tfrac{1}{2} - 2 \cdot 2^{-2r} > \tfrac{1}{4}$$

(where the last inequality holds because we assume that k', ℓ', r' and thus r are sufficiently large.) Thus,

$$\left\|\mathbb{P}_{x|e}\right\|_2 \leq c_e^{-1} \cdot \left\|\mathbb{P}_{x|v}\right\|_2 \leq 4 \cdot 2^{\ell} \cdot 2^{-n}$$ □

Claim 4.7.

$$\left\|\mathbb{P}_{x|s}\right\|_2 \leq 4 \cdot 2^{\ell} \cdot 2^{-n}.$$

Proof. Let $\Gamma_{in}(s)$ be the set of all edges e of B, that are going into s, such that $\Pr(e) > 0$. Note that

$$\sum_{e \in \Gamma_{in}(s)} \Pr(e) = \Pr(s).$$

By the law of total probability, for every $x' \in X$,

$$\mathbb{P}_{x|s}(x') = \sum_{e \in \Gamma_{in}(s)} \frac{\Pr(e)}{\Pr(s)} \cdot \mathbb{P}_{x|e}(x'),$$

and hence by Jensen's inequality,

$$\mathbb{P}_{x|s}(x')^2 \leq \sum_{e \in \Gamma_{in}(s)} \frac{\Pr(e)}{\Pr(s)} \cdot \mathbb{P}_{x|e}(x')^2.$$

Summing over $x' \in X$, we obtain,

$$\left\|\mathbb{P}_{x|s}\right\|_2^2 \leq \sum_{e \in \Gamma_{in}(s)} \frac{\Pr(e)}{\Pr(s)} \cdot \left\|\mathbb{P}_{x|e}\right\|_2^2.$$

By Claim 4.6, for any $e \in \Gamma_{in}(s)$,

$$\left\|\mathbb{P}_{x|e}\right\|_2^2 \leq \left(4 \cdot 2^{\ell} \cdot 2^{-n}\right)^2.$$

Hence,

$$\left\|\mathbb{P}_{x|s}\right\|_2^2 \leq \left(4 \cdot 2^{\ell} \cdot 2^{-n}\right)^2. \qquad \square$$

Similarity to a Target Distribution

Recall that for two functions $f, g : X \to \mathbb{R}^+$, we defined

$$\langle f, g \rangle = \mathop{\mathbf{E}}_{z \in_R X}[f(z) \cdot g(z)].$$

We think of $\langle f, g \rangle$ as a measure for the similarity between a function f and a target function g. Typically f, g will be distributions.

Claim 4.8.

$$\langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle > 2^{2\ell} \cdot 2^{-2n}.$$

Proof. Since s is significant,

$$\langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle = \left\|\mathbb{P}_{x|s}\right\|_2^2 > 2^{2\ell} \cdot 2^{-2n}. \qquad \square$$

Claim 4.9.

$$\langle \mathcal{U}_X, \mathbb{P}_{x|s} \rangle = 2^{-2n},$$

where $\mathcal{U}_X$ is the uniform distribution over X.

Proof. Since $\mathbb{P}_{x|s}$ is a distribution,

$$\langle \mathcal{U}_X, \mathbb{P}_{x|s} \rangle = 2^{-2n} \cdot \sum_{z \in X} \mathbb{P}_{x|s}(z) = 2^{-2n}. \qquad \square$$

Measuring the Progress

For $i \in \{0, \ldots, m\}$, let L_i be the set of vertices v in layer-i of B, such that $\Pr(v) > 0$. For $i \in \{1, \ldots, m\}$, let Γ_i be the set of edges e from layer-$(i-1)$ of B to layer-i of B, such that $\Pr(e) > 0$. Recall that $k = \gamma k'$ (Equation (2)).

For $i \in \{0, \ldots, m\}$, let

$$\mathcal{Z}_i = \sum_{v \in L_i} \Pr(v) \cdot \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^k.$$

For $i \in \{1, \ldots, m\}$, let

$$\mathcal{Z}'_i = \sum_{e \in \Gamma_i} \Pr(e) \cdot \langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^k.$$

We think of $\mathcal{Z}_i, \mathcal{Z}'_i$ as measuring the progress made by the branching program, towards reaching a state with distribution similar to $\mathbb{P}_{x|s}$.

For a vertex v of B, let $\Gamma_{out}(v)$ be the set of all edges e of B, that are going out of v, such that $\Pr(e) > 0$. Note that

$$\sum_{e \in \Gamma_{out}(v)} \Pr(e) \leq \Pr(v).$$

(We don't always have an equality here, since sometimes $\mathcal{T}$ stops on v).

The next four claims show that the progress made by the branching program is slow.

Claim 4.10. *For every vertex v of B, such that $\Pr(v) > 0$,*

$$\sum_{e \in \Gamma_{out}(v)} \frac{\Pr(e)}{\Pr(v)} \cdot \langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^k \leq \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^k \cdot \left(1 + 2^{-r}\right)^k + \left(2^{-2n+2}\right)^k.$$

Proof. If v is significant or v is a leaf, then $\mathcal{T}$ always stops on v and hence $\Gamma_{out}(v)$ is empty and thus the left hand side is equal to zero and the right hand side is positive, so the claim follows trivially. Thus, we can assume that v is not significant and is not a leaf.

Define $P : X \to \mathbb{R}^+$ as follows. For any $x' \in X$,

$$P(x') = \begin{cases} 0 & \text{if} \quad x' \in \mathrm{Sig}(v) \\ \mathbb{P}_{x|v}(x') & \text{if} \quad x' \notin \mathrm{Sig}(v) \end{cases}$$

Note that by the definition of $\mathrm{Sig}(v)$, for any $x' \in X$,

$$P(x') \leq 2^{2\ell+2r} \cdot 2^{-n}. \tag{4}$$

Define $f : X \to \mathbb{R}^+$ as follows. For any $x' \in X$,

$$f(x') = P(x') \cdot \mathbb{P}_{x|s}(x').$$

By Claim 4.7 and Equation (4),

$$\|f\|_2 \leq 2^{2\ell+2r} \cdot 2^{-n} \cdot \left\|\mathbb{P}_{x|s}\right\|_2 \leq 2^{2\ell+2r} \cdot 2^{-n} \cdot 4 \cdot 2^{\ell} \cdot 2^{-n} = 2^{3\ell+2r+2} \cdot 2^{-2n}. \tag{5}$$

By Claim 4.5, for any edge $e \in \Gamma_{out}(v)$, labeled by (a, b), for any $x' \in X$,

$$\mathbb{P}_{x|e}(x') = \begin{cases} 0 & \text{if} \quad M(a, x') \neq b \\ P(x') \cdot c_e^{-1} & \text{if} \quad M(a, x') = b \end{cases}$$

where c_e satisfies,

$$c_e \geq \tfrac{1}{2} - 2 \cdot 2^{-2r}.$$

Therefore, for any edge $e \in \Gamma_{out}(v)$, labeled by (a,b), for any $x' \in X$,

$$\mathbb{P}_{x|e}(x') \cdot \mathbb{P}_{x|s}(x') = \begin{cases} 0 & \text{if} \quad M(a,x') \neq b \\ f(x') \cdot c_e^{-1} & \text{if} \quad M(a,x') = b \end{cases}$$

and hence, we have

$$\begin{aligned} \langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle &= \mathop{\mathbf{E}}_{x' \in_R X}[\mathbb{P}_{x|e}(x') \cdot \mathbb{P}_{x|s}(x')] = \mathop{\mathbf{E}}_{x' \in_R X}[f(x') \cdot c_e^{-1} \cdot \mathbf{1}_{\{x' \in X \,:\, M(a,x')=b\}}] \\ &= \mathop{\mathbf{E}}_{x' \in_R X}\left[f(x') \cdot c_e^{-1} \cdot \tfrac{(1+b \cdot M(a,x'))}{2}\right] = (\|f\|_1 + b \cdot \langle M_a, f \rangle) \cdot (2c_e)^{-1} \\ &< (\|f\|_1 + |\langle M_a, f \rangle|) \cdot \left(1 + 2^{-2r+3}\right) \end{aligned} \tag{6}$$

(where the last inequality holds by the bound that we have on c_e, because we assume that k', ℓ', r' and thus r are sufficiently large).

We will now consider two cases:

Case I: $\|f\|_1 < 2^{-2n}$

In this case, we bound $|\langle M_a, f \rangle| \leq \|f\|_1$ (since f is non-negative and the entries of M are in $\{-1,1\}$) and $(1 + 2^{-2r+3}) < 2$ (since we assume that k', ℓ', r' and thus r are sufficiently large) and obtain for any edge $e \in \Gamma_{out}(v)$,

$$\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle < 4 \cdot 2^{-2n}.$$

Since $\sum_{e \in \Gamma_{out}(v)} \frac{\Pr(e)}{\Pr(v)} \leq 1$, Claim 4.10 follows, as the left hand side of the claim is smaller than the second term on the right hand side.

Case II: $\|f\|_1 \geq 2^{-2n}$

For every $a \in A$, define

$$t(a) = \frac{|\langle M_a, f \rangle|}{\|f\|_1}.$$

By Equation (6),

$$\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^k < \|f\|_1^k \cdot (1 + t(a))^k \cdot \left(1 + 2^{-2r+3}\right)^k. \tag{7}$$

Note that by the definitions of P and f,

$$\|f\|_1 = \mathop{\mathbf{E}}_{x' \in_R X}[f(x')] = \langle P, \mathbb{P}_{x|s} \rangle \leq \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle.$$

Note also that for every $a \in A$, there is at most one edge $e_{(a,1)} \in \Gamma_{out}(v)$, labeled by $(a,1)$, and at most one edge $e_{(a,-1)} \in \Gamma_{out}(v)$, labeled by $(a,-1)$, and we have

$$\tfrac{\Pr(e_{(a,1)})}{\Pr(v)} + \tfrac{\Pr(e_{(a,-1)})}{\Pr(v)} \leq \tfrac{1}{|A|},$$

since $\frac{1}{|A|}$ is the probability that the next sample read by the program is a. Thus, summing over all $e \in \Gamma_{out}(v)$, by Equation (7),

$$\sum_{e \in \Gamma_{out}(v)} \frac{\Pr(e)}{\Pr(v)} \cdot \langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^k < \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^k \cdot \mathop{\mathbf{E}}_{a \in_R A} \left[(1 + t(a))^k \right] \cdot \left(1 + 2^{-2r+3}\right)^k . \tag{8}$$

It remains to bound

$$\mathop{\mathbf{E}}_{a \in_R A} \left[(1 + t(a))^k \right], \tag{9}$$

using the properties of the matrix M and the bounds on the ℓ_2 versus ℓ_1 norms of f.

By Equation (5), the assumption that $\|f\|_1 \geq 2^{-2n}$, Equation (1) and Equation (2), we get

$$\frac{\|f\|_2}{\|f\|_1} \leq 2^{3\ell+2r+2} \leq 2^{\ell'} .$$

Since M is a (k', ℓ')-L_2-extractor with error $2^{-r'}$, there are at most $2^{-k'} \cdot |A|$ rows $a \in A$ with $t(a) = \frac{|\langle M_a, f \rangle|}{\|f\|_1} \geq 2^{-r'}$. We bound the expectation in Equation (9), by splitting the expectation into two sums

$$\mathop{\mathbf{E}}_{a \in_R A} \left[(1 + t(a))^k \right] = \tfrac{1}{|A|} \cdot \sum_{a \,:\, t(a) \leq 2^{-r'}} (1 + t(a))^k + \tfrac{1}{|A|} \cdot \sum_{a \,:\, t(a) > 2^{-r'}} (1 + t(a))^k . \tag{10}$$

We bound the first sum in Equation (10) by $(1 + 2^{-r'})^k$. As for the second sum in Equation (10), we know that it is a sum of at most $2^{-k'} \cdot |A|$ elements, and since for every $a \in A$, we have $t(a) \leq 1$, we have

$$\tfrac{1}{|A|} \cdot \sum_{a \,:\, t(a) > 2^{-r'}} (1 + t(a))^k \leq 2^{-k'} \cdot 2^k \leq 2^{-2r}$$

(where in the last inequality we used Equations (1) and (2)). Overall, using Equation (1) again, we get

$$\mathop{\mathbf{E}}_{a \in_R A} \left[(1 + t(a))^k \right] \leq (1 + 2^{-r'})^k + 2^{-2r} \leq (1 + 2^{-2r})^{k+1}. \tag{11}$$

Substituting Equation (11) into Equation (8), we obtain

$$\begin{aligned} \sum_{e \in \Gamma_{out}(v)} \frac{\Pr(e)}{\Pr(v)} \cdot \langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^k &< \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^k \cdot \left(1 + 2^{-2r}\right)^{k+1} \cdot \left(1 + 2^{-2r+3}\right)^k \\ &< \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^k \cdot \left(1 + 2^{-r}\right)^k \end{aligned}$$

(where the last inequality uses the assumption that r is sufficiently large). This completes the proof of Claim 4.10. □

Claim 4.11. *For every $i \in \{1, \ldots, m\}$,*

$$\mathcal{Z}'_i \leq \mathcal{Z}_{i-1} \cdot \left(1 + 2^{-r}\right)^k + \left(2^{-2n+2}\right)^k .$$

Proof. By Claim 4.10,

$$\begin{aligned}\mathcal{Z}'_i = \sum_{e\in\Gamma_i} \Pr(e)\cdot\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s}\rangle^k &= \sum_{v\in L_{i-1}} \Pr(v)\cdot \sum_{e\in\Gamma_{out}(v)} \tfrac{\Pr(e)}{\Pr(v)}\cdot\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s}\rangle^k \\ &\le \sum_{v\in L_{i-1}} \Pr(v)\cdot\left(\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s}\rangle^k\cdot\left(1+2^{-r}\right)^k + \left(2^{-2n+2}\right)^k\right) \\ &= \mathcal{Z}_{i-1}\cdot\left(1+2^{-r}\right)^k + \sum_{v\in L_{i-1}} \Pr(v)\cdot\left(2^{-2n+2}\right)^k \\ &\le \mathcal{Z}_{i-1}\cdot\left(1+2^{-r}\right)^k + \left(2^{-2n+2}\right)^k \end{aligned}$$

□

Claim 4.12. *For every $i\in\{1,\ldots,m\}$,*

$$\mathcal{Z}_i \le \mathcal{Z}'_i.$$

Proof. For any $v\in L_i$, let $\Gamma_{in}(v)$ be the set of all edges $e\in\Gamma_i$, that are going into v. Note that

$$\sum_{e\in\Gamma_{in}(v)} \Pr(e) = \Pr(v).$$

By the law of total probability, for every $v\in L_i$ and every $x'\in X$,

$$\mathbb{P}_{x|v}(x') = \sum_{e\in\Gamma_{in}(v)} \tfrac{\Pr(e)}{\Pr(v)}\cdot\mathbb{P}_{x|e}(x'),$$

and hence

$$\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s}\rangle = \sum_{e\in\Gamma_{in}(v)} \tfrac{\Pr(e)}{\Pr(v)}\cdot\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s}\rangle.$$

Thus, by Jensen's inequality,

$$\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s}\rangle^k \le \sum_{e\in\Gamma_{in}(v)} \tfrac{\Pr(e)}{\Pr(v)}\cdot\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s}\rangle^k.$$

Summing over all $v\in L_i$, we get

$$\begin{aligned}\mathcal{Z}_i = \sum_{v\in L_i} \Pr(v)\cdot\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s}\rangle^k &\le \sum_{v\in L_i} \Pr(v)\cdot \sum_{e\in\Gamma_{in}(v)} \tfrac{\Pr(e)}{\Pr(v)}\cdot\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s}\rangle^k \\ &= \sum_{e\in\Gamma_i} \Pr(e)\cdot\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s}\rangle^k = \mathcal{Z}'_i.\end{aligned}$$

□

Claim 4.13. *For every $i\in\{1,\ldots,m\}$,*

$$\mathcal{Z}_i \le 2^{4k+2r}\cdot 2^{-2k\cdot n}.$$

Proof. By Claim 4.9, $\mathcal{Z}_0 = (2^{-2n})^k$. By Claim 4.11 and Claim 4.12, for every $i\in\{1,\ldots,m\}$,

$$\mathcal{Z}_i \le \mathcal{Z}_{i-1}\cdot\left(1+2^{-r}\right)^k + \left(2^{-2n+2}\right)^k.$$

Hence, for every $i\in\{1,\ldots,m\}$,

$$\mathcal{Z}_i \le \left(2^{-2n+2}\right)^k\cdot m\cdot\left(1+2^{-r}\right)^{km}.$$

Since $m = 2^r$,

$$\mathcal{Z}_i \le 2^{-2k\cdot n}\cdot 2^{2k}\cdot 2^r\cdot e^k \le 2^{-2k\cdot n}\cdot 2^{4k+2r}.$$

□

Proof of Lemma 4.1

We can now complete the proof of Lemma 4.1. Assume that s is in layer-i of B. By Claim 4.8,

$$\mathcal{Z}_i \geq \Pr(s) \cdot \langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle^k > \Pr(s) \cdot \left(2^{2\ell} \cdot 2^{-2n}\right)^k = \Pr(s) \cdot 2^{2\ell \cdot k} \cdot 2^{-2k \cdot n}.$$

On the other hand, by Claim 4.13,

$$\mathcal{Z}_i \leq 2^{4k+2r} \cdot 2^{-2k \cdot n}.$$

Thus, using Equation (1) and Equation (2), we get

$$\Pr(s) \leq 2^{4k+2r} \cdot 2^{-2\ell \cdot k} \leq 2^{4k'} \cdot 2^{-(2\gamma^2/3) \cdot (k'\ell')}.$$

Recall that we assumed that the width of B is at most $2^{ck'\ell'}$ for some constant $c < 2/3$, and that the length of B is at most 2^r. Recall that we fixed γ such that $2\gamma^2/3 > c$. Taking a union bound over at most $2^r \cdot 2^{ck'\ell'} \leq 2^{k'} \cdot 2^{ck'\ell'}$ significant vertices of B, we conclude that the probability that $\mathcal{T}$ reaches any significant vertex is at most $2^{-\Omega(k'\ell')}$. Since we assume that k' and ℓ' are sufficiently large, $2^{-\Omega(k'\ell')}$ is certainly at most $2^{-k'}$, which is at most 2^{-r}. □

4.4 Lower Bounds for Weak Learning

In this section, we show that under the same conditions of Theorem 1, the branching program cannot even weakly-learn the function. That is, we show that the branching program cannot output a hypothesis $h : A \to \{-1, 1\}$ with a non-negligible correlation with the function defined by the true unknown x. We change the definition of the branching program and associate with each leaf v a hypothesis $h_v : A \to \{-1, 1\}$. We measure the success as the correlation between h_v and the function defined by the true unknown x.

Formally, for any $x \in X$, let $M^{(x)} : A \to \{-1, 1\}$ be the function corresponding to the x-th column of M. We define the value of the program as $\mathbf{E}\left[\left|\langle h_v, M^{(x)} \rangle\right|\right]$, where the expectation is over $x, a_1, \ldots, a_m$ (recall that x is uniformly distributed over X and $a_1, \ldots, a_m$ are uniformly distributed over A, and for every t, $b_t = M(a_t, x)$). The following claim bounds the expected correlation between h_v and $M^{(x)}$, conditioned on reaching a non-significant leaf.

Claim 4.14. *If v is a non-significant leaf, then*

$$\mathop{\mathbf{E}}_{x}\left[\left|\langle h_v, M^{(x)} \rangle\right| \;\middle|\; E_v\right] \leq O(2^{-r/2}).$$

Proof. We expand the expected correlation between h_v and $M^{(x)}$, squared:

$$\begin{aligned}
\mathop{\mathbf{E}}_{x}\left[\left|\langle h_v, M^{(x)} \rangle\right| \;\middle|\; E_v\right]^2 &\leq \mathop{\mathbf{E}}_{x}\left[\langle h_v, M^{(x)} \rangle^2 \;\middle|\; E_v\right] = \sum_{x' \in X} \mathbb{P}_{x|v}(x') \cdot \langle h_v, M^{(x')} \rangle^2 \\
&= \sum_{x' \in X} \mathbb{P}_{x|v}(x') \cdot \mathop{\mathbf{E}}_{a, a' \in_R A}[h_v(a) \cdot M(a, x') \cdot h_v(a') \cdot M(a', x')] \\
&= \mathop{\mathbf{E}}_{a, a' \in_R A}\left[h_v(a) \cdot h_v(a') \cdot \sum_{x' \in X} \mathbb{P}_{x|v}(x') \cdot M(a, x') \cdot M(a', x')\right] \\
&\leq \mathop{\mathbf{E}}_{a, a' \in_R A}\left[\left|\sum_{x' \in X} \mathbb{P}_{x|v}(x') \cdot M(a, x') \cdot M(a', x')\right|\right] \\
&= \mathop{\mathbf{E}}_{a \in_R A}\left[\mathop{\mathbf{E}}_{a' \in_R A}\left[\left|\sum_{x' \in X} \mathbb{P}_{x|v}(x') \cdot M(a, x') \cdot M(a', x')\right|\right]\right] .
\end{aligned}$$

Next, we show that $\mathbf{E}_{a' \in_R A}\left[\left|\sum_{x' \in X} \mathbb{P}_{x|v}(x') \cdot M(a,x') \cdot M(a',x')\right|\right] \leq 4 \cdot 2^{-r}$ for any $a \in A$. Fix $a \in A$. Let $q_a : X \to \mathbb{R}$ be the function defined by $q_a(x') = \mathbb{P}_{x|v}(x') \cdot M(a,x')$ for $x' \in X$. Since $|q_a(x')| = |\mathbb{P}_{x|v}(x')|$ for any $x' \in X$ and since v is a non-significant vertex, we get

$$\|q_a\|_2 = \left\|\mathbb{P}_{x|v}\right\|_2 \leq 2^{\ell} \cdot 2^{-n} \qquad \text{and} \qquad \|q_a\|_1 = \left\|\mathbb{P}_{x|v}\right\|_1 = 2^{-n}.$$

Hence, $\frac{\|q_a\|_2}{\|q_a\|_1} \leq 2^{\ell}$. We would like to use the fact that M is a (k',ℓ')-L_2-extractor with error $2^{-r'}$ to show that there aren't many rows of M with a large inner product with q_a. However, q_a can get negative values and the definition of L_2-extractors only handles non-negative functions $f : X \to \mathbb{R}^+$. To solve this issue, we use the following lemma, proved in Section 5.1.

Lemma 4.15. *Suppose that $M : A \times X \to \{-1,1\}$ is a (k',ℓ')-L_2-extractor with error at most 2^{-r}. Let $f : X \to \mathbb{R}$ be any function (i.e., f can get negative values) with $\frac{\|f\|_2}{\|f\|_1} \leq 2^{\ell'-r}$. Then, there are at most $2 \cdot 2^{-k'} \cdot |A|$ rows $a \in A$ with $\frac{|\langle M_a, f\rangle|}{\|f\|_1} \geq 2 \cdot 2^{-r}$.*

Since M is a (k',ℓ')-L_2-extractor with error at most $2^{-r'}$, and since $r < r'$, we have that M is also a (k',ℓ')-L_2-extractor with error at most 2^{-r}. Since $\frac{\|q_a\|_2}{\|q_a\|_1} \leq 2^{\ell} \leq 2^{\ell'-r}$, we can apply Lemma 4.15 with $f = q_a$, and error 2^{-r}. We get that there are at most $2 \cdot 2^{-k'} \cdot |A|$ rows $a' \in A$ with $\frac{|\langle q_a, M_{a'}\rangle|}{\|q_a\|_1} \geq 2 \cdot 2^{-r}$. Thus,

$$\mathop{\mathbf{E}}_{a' \in_R A}\left[\left|\sum_{x' \in X} q_a(x') \cdot M(a',x')\right|\right] = \mathop{\mathbf{E}}_{a' \in_R A}\left[\frac{|\langle q_a, M_{a'}\rangle|}{\|q_a\|_1}\right] \leq 2 \cdot 2^{-k'} + 2 \cdot 2^{-r} \leq 4 \cdot 2^{-r} .$$

Overall, we get that $\mathbf{E}_x\left[|\langle h_v, M^{(x)}\rangle| \;\middle|\; E_v\right]^2 \leq 4 \cdot 2^{-r}$. Taking square roots of both sides of the last inequality completes the proof. □

Lemma 4.1, Claim 4.2 and Claim 4.3 show that the probability that $\mathcal{T}$ stops before reaching a leaf is at most $O(2^{-r})$. Combining this with Claim 4.14 we get that (under the same conditions of Theorem 1)

$$\mathbf{E}[\left|\langle h_v, M^{(x)}\rangle\right|] \leq \Pr[\mathcal{T} \text{ stops}] + O(2^{-r/2}) \leq O(2^{-r/2}),$$

where the expectation and probability are taken over $x \in_R X$ and $a_1, \ldots, a_m \in_R A$. We get the following theorem as a conclusion.

Theorem 2. *Let $\frac{1}{100} < c < \frac{2}{3}$. Fix γ to be such that $\frac{3c}{2} < \gamma^2 < 1$.*

Let X, A be two finite sets. Let $n = \log_2 |X|$. Let $M : A \times X \to \{-1,1\}$ be a matrix which is a (k',ℓ')-L_2-extractor with error $2^{-r'}$, for sufficiently large[2] k', ℓ' and r', where $\ell' \leq n$. Let

$$r := \min\left\{\tfrac{r'}{2}, \tfrac{(1-\gamma)k'}{2}, \tfrac{(1-\gamma)\ell'}{2} - 1\right\}.$$

Let B be a branching program of length at most 2^r and width at most $2^{c \cdot k' \cdot \ell'}$ for the learning problem that corresponds to the matrix M. Then,

$$\mathbf{E}[\left|\langle h_v, M^{(x)}\rangle\right|] \leq O(2^{-r/2}) .$$

In particular, the probability that the hypothesis agrees with the function defined by the true unknown x, on more than $1/2 + 2^{-r/4}$ of the inputs, is at most $O(2^{-r/4})$.

[2] By *"sufficiently large"* we mean that k', ℓ', r' are larger than some constant that depends on γ.

4.5 Main Corollary

Corollary 3. *There exists a (sufficiently small) constant $c > 0$, such that:*

Let X, A be two finite sets. Let $M : A \times X \to \{-1, 1\}$ be a matrix. Assume that $k, \ell, r \in \mathbb{N}$ are such that any submatrix of M of at least $2^{-k} \cdot |A|$ rows and at least $2^{-\ell} \cdot |X|$ columns, has a bias of at most 2^{-r}.

Let B be a branching program of length at most $2^{c \cdot r}$ and width at most $2^{c \cdot k \cdot \ell}$ for the learning problem that corresponds to the matrix M. Then, the success probability of B is at most $2^{-\Omega(r)}$.

Proof. By Lemma 5.2 (stated and proved below), there exist $k' = k + \Omega(r)$, $\ell' = \ell + \Omega(r)$, and $r' = \Omega(r)$, such that: any submatrix of M of at least $2^{-k'} \cdot |A|$ rows and at least $2^{-\ell'} \cdot |X|$ columns, has a bias of at most $2^{-r'}$.

By Lemma 5.4 (stated and proved below), M is an $(\Omega(k)+\Omega(r), \Omega(\ell)+\Omega(r))$-$L_2$-extractor with error $2^{-\Omega(r)}$.

The corollary follows by Theorem 1. □

5 Applications

5.1 Some Useful Lemmas

5.1.1 Handling Negative Functions

In the following lemma, we show that up to a small loss in parameters an L_2-extractor has similar guarantees for any function $f : X \to \mathbb{R}$ with bounded ℓ_2-vs-ℓ_1-norm regardless of whether or not f is non-negative.

Lemma 5.1. *Suppose that $M : A \times X \to \{-1, 1\}$ is a (k', ℓ')-L_2-extractor with error at most 2^{-r}. Let $f : X \to \mathbb{R}$ be any function with $\frac{\|f\|_2}{\|f\|_1} \leq 2^{\ell'-r}$. Then, there are at most $2 \cdot 2^{-k'} \cdot |A|$ rows $a \in A$ with $\frac{|\langle M_a, f \rangle|}{\|f\|_1} \geq 2 \cdot 2^{-r}$.*

Proof. Let $f_+, f_- : X \to \mathbb{R}^+$ be the non-negative functions defined by

$$f_+(x) = \begin{cases} f(x), & f(x) > 0 \\ 0, & \text{otherwise} \end{cases} \qquad f_-(x) = \begin{cases} |f(x)|, & f(x) < 0 \\ 0, & \text{otherwise} \end{cases}$$

for $x \in X$. We have $f(x) = f_+(x) - f_-(x)$ for all $x \in X$. We split into two cases:

1. If $\|f_+\|_1 < 2^{-r} \cdot \|f\|_1$, then $|\langle M_a, f_+ \rangle| \leq \|f_+\|_1 < 2^{-r} \cdot \|f\|_1$ for all $a \in A$.
2. If $\|f_+\|_1 \geq 2^{-r} \cdot \|f\|_1$, then f_+ is a non-negative function with

$$\frac{\|f_+\|_2}{\|f_+\|_1} \leq \frac{\|f\|_2}{\|f\|_1 \cdot 2^{-r}} \leq 2^{\ell'} .$$

Thus, we may use the assumption that M is an L_2-extractor to deduce that there are at most $2^{-k'} \cdot |A|$ rows $a \in A$ with $|\langle M_a, f_+ \rangle| \geq \|f_+\|_1 \cdot 2^{-r}$.

In both cases, there are at most $2^{-k'} \cdot |A|$ rows $a \in A$ with $|\langle M_a, f_+ \rangle| \geq \|f\|_1 \cdot 2^{-r}$. Similarly, there are at most $2^{-k'} \cdot |A|$ rows $a \in A$ with $|\langle M_a, f_- \rangle| \geq \|f\|_1 \cdot 2^{-r}$. Thus, for all but at most $2 \cdot 2^{-k'} \cdot |A|$ of the rows $a \in A$ we have

$$|\langle M_a, f \rangle| \leq |\langle M_a, f_+ \rangle| + |\langle M_a, f_- \rangle| < 2 \cdot \|f\|_1 \cdot 2^{-r} . \qquad \square$$

5.1.2 Error vs. Min-Entropy

Lemma 5.2. *Let $M : A \times X \to \{-1, 1\}$ be a matrix. Let k, ℓ, r be such that any submatrix of M of at least $2^{-k} \cdot |A|$ rows and at least $2^{-\ell} \cdot |X|$ columns, has a bias of at most 2^{-r}.*

Then, there exist $k' = k + \Omega(r)$, $\ell' = \ell + \Omega(r)$, and $r' = \Omega(r)$, such that: any submatrix of M of at least $2^{-k'} \cdot |A|$ rows and at least $2^{-\ell'} \cdot |X|$ columns, has a bias of at most $2^{-r'}$.

Proof. Assume without loss of generality that k, ℓ, r are larger than some sufficiently large absolute constant.

We will show that there exists $k' = k + \Omega(r)$, such that, any submatrix of M of at least $2^{-k'} \cdot |A|$ rows and at least $2^{-\ell} \cdot |X|$ columns, has a bias of at most $2^{-\Omega(r)}$. The proof of the lemma then follows by applying the same claim again on the transposed matrix.

Let $k' = k + \frac{r}{10}$. Assume for a contradiction that there exist $T \subseteq A$ of size at least $2^{-k'} \cdot |A|$ and $S \subseteq X$ of size at least $2^{-\ell} \cdot |X|$, such that the bias of $T \times S$ is larger than, say, $2^{-r/2}$. By the assumption of the lemma, $|T| < 2^{-k} \cdot |A|$.

Let T' be an arbitrary set of $2^{-k} \cdot |A|$ rows in $A \setminus T$. By the assumption of the lemma, the bias of $T' \times S$ is at most 2^{-r}. Therefore, the bias of $(T' \cup T) \times S$ is at least

$$\tfrac{|T|}{|T' \cup T|} \cdot 2^{-r/2} - \tfrac{|T'|}{|T' \cup T|} \cdot 2^{-r} \geq \tfrac{1}{2} \cdot 2^{-r/10} \cdot 2^{-r/2} - 2^{-r} > 2^{-r}.$$

Thus, $(T' \cup T) \times S$ contradicts the assumption of the lemma. $\square$

5.1.3 L_2-Extractors and L_∞-Extractors

We will show that M being an L_2-Extractor is equivalent to M being an L_∞-Extractor (barring constants).

Lemma 5.3. *If a matrix $M : A \times X \to \{-1, 1\}$ is a (k, ℓ)-L_2-Extractor with error 2^{-r}, then M is also a $(k - \xi, 2\ell \sim (\min\{r, \xi\} - 1))$-$L_\infty$-Extractor, $\forall 0 < \xi < k$.*

Taking $\xi = \frac{k}{2}$, we get that if M is a (k, ℓ)-L_2-Extractor with error 2^{-r}, then M is also a $(\Omega(k), \Omega(\ell) \sim (\Omega(\min\{r, k\})))$-$L_\infty$-Extractor.

Proof. We pick a ξ $(0 < \xi < k)$. To prove that M is a $(k - \xi, 2\ell \sim (\min\{r, \xi\} - 1))$-$L_\infty$-Extractor, it suffices to prove the statement of the L_∞-Extractors for any two uniform distributions over subsets $A_1 \subseteq A$ and $X_1 \subseteq X$ of size at least $\frac{|A|}{2^{k-\xi}}$ and $\frac{|X|}{2^{2\ell}}$ respectively. This follows from the fact that any distribution with min-entropy at least h can be written as a convex combination of uniform distributions on sets of size at least 2^h [CG88].

For a distribution p_x, which is uniform over a subset $X_1 \subseteq X$ of size at least $\frac{|X|}{2^{2\ell}}$,

$$\frac{\|p_x\|_2}{\|p_x\|_1} = \left(\frac{|X|}{|X_1|}\right)^{\frac{1}{2}} \leq 2^{\ell}.$$

Using the fact that M is a (k,ℓ)-L_2-Extractor with error 2^{-r}, we know that there are at most $\frac{|A|}{2^k}$ rows a with $|(M \cdot p_x)_a| \geq 2^{-r}$. Using the fact that p_a is a uniform distribution over a set A_1 of size at least $\frac{|A|}{2^{k-\xi}}$, we get

$$\begin{aligned}\left|\sum_{a' \in A}\sum_{x' \in X} p_a(a') \cdot p_x(x') \cdot M(a',x')\right| &\leq \frac{1}{|A_1|} \cdot \sum_{a' \in A_1} |(M \cdot p_x)_{a'}| \\ &\leq \frac{1}{|A_1|} \cdot \left(\frac{|A|}{2^k} + |A_1| \cdot 2^{-r}\right) \leq 2^{-\xi} + 2^{-r}\end{aligned}$$

This proves that M is a $(k-\xi, 2\ell \sim (\min\{r,\xi\}-1))$-$L_\infty$-Extractor, $\forall 0 < \xi < k$. □

Lemma 5.4. *If a matrix $M : A \times X \to \{-1,1\}$ is a $(k, \ell \sim r)$-L_∞-Extractor, then M is also a $\left(k-1, \frac{\ell-\xi-1}{2}\right)$-$L_2$-Extractor with error $2^{-r} + 2^{-\xi+1}$, $\forall 1 \leq \xi \leq \ell - 1$.*

Taking $\xi = \frac{\ell}{2}$, we get that if M is a $(k, \ell \sim r)$-L_∞-Extractor, then M is also a $(\Omega(k), \Omega(\ell))$-$L_2$-Extractor with error $2^{-\Omega(\min\{r,\ell\})}$.

In this proof, we use the following notation. For two non-negative functions $P, Q : X \to \mathbb{R}$, we denote by $\text{dist}(P,Q)$ the ℓ_1-distance between the two functions, that is

$$\text{dist}(P,Q) = \sum_{x \in X} |P(x) - Q(x)| \,.$$

Note that $\text{dist}(P,Q) = \|P - Q\|_1 \cdot |X|$.

Proof. We want to prove that for any $1 \leq \xi \leq \ell - 1$, and any non-negative function $f : X \to \mathbb{R}$ with $\frac{\|f\|_2}{\|f\|_1} \leq 2^{\frac{\ell-\xi-1}{2}}$, there are at most $2 \cdot 2^{-k} \cdot |A|$ rows $a \in A$ with $\frac{|\langle M_a, f\rangle|}{\|f\|_1} \geq 2^{-r} + 2^{-\xi+1}$.

Let's assume that there exists a non-negative function $f : X \to \mathbb{R}$ for which the last statement is not true. Let f_p be a probability distribution on X defined by $f_p(x) = \frac{f(x)}{\sum_x f(x)} = \frac{f(x)}{|X| \cdot \|f\|_1}$. Then,

$$\begin{aligned}\|f_p\|_2 = \frac{\|f\|_2}{|X| \cdot \|f\|_1} &\leq \frac{2^{\frac{\ell-\xi-1}{2}}}{|X|} \\ \implies \left(\frac{\sum_x f_p(x)^2}{|X|}\right)^{\frac{1}{2}} &\leq \frac{2^{\frac{\ell-\xi-1}{2}}}{|X|} \\ \implies \sum_x f_p(x)^2 &\leq 2^{\ell-\xi-1-\log(|X|)}\end{aligned}$$

Thus, there is strictly less than $2^{-\xi}$ probability mass on elements x with $f_p(x) > 2^{\ell-\log(|X|)-1}$. Let $\bar{f}_p : X \to \mathbb{R}$ be the trimmed function that takes values $f_p(x)$ at x when $f_p(x) \leq 2^{\ell-\log(|X|)-1}$ and 0 otherwise. We define a new probability distribution $p_x : X \to [0,1]$ as

$$p_x(x') = \bar{f}_p(x') + \frac{1 - \sum_{x'} \bar{f}_p(x')}{|X|}.$$

Informally, we are just redistributing the probability mass removed from f_p. It is easy to see that the new probability distribution p_x has min-entropy at least $\log(|X|) - \ell$, and

$$\text{dist}(p_x, f_p) < 2^{-\xi+1} \tag{12}$$

as $\text{dist}(p_x, f_p) \leq \text{dist}(p_x, \bar{f}_p) + \text{dist}(\bar{f}_p, f_p) < 2^{-\xi} + 2^{-\xi}$.

Let A_{bad} be the set of rows $a \in A$ with $\frac{|\langle M_a, f\rangle|}{\|f\|_1} = |(M \cdot f_p)_a| \geq 2^{-r} + 2^{-\xi+1}$. By our assumption, $|A_{\text{bad}}| \geq 2 \cdot 2^{-k}|A|$. Let A_1 and A_2 be the set of rows a with $(M \cdot f_p)_a \geq 2^{-r} + 2^{-\xi+1}$ and $(M \cdot f_p)_a \leq -(2^{-r} + 2^{-\xi+1})$ respectively. As $A_{\text{bad}} = A_1 \cup A_2$, w.l.o.g. $|A_1| \geq |A_{\text{bad}}|/2 \geq 2^{-k}|A|$ (else we can work with A_2 and the rest of the argument follows similarly). Let p_a be a uniform probability distribution over the set A_1. Clearly p_a has min-entropy at least $\log(|A|) - k$.

As $(M \cdot f_p)_a \geq 2^{-r} + 2^{-\xi+1}$ for the entire support of p_a, we get

$$\left| \mathop{\mathbf{E}}_{a \in_R A_1} [(M \cdot f_p)_a] \right| \geq 2^{-r} + 2^{-\xi+1}. \tag{13}$$

As the entries of M have magnitude at most 1, we have

$$\left| \mathop{\mathbf{E}}_{a \in_R A_1} [(M \cdot (p_x - f_p))_a] \right| \leq \mathop{\mathbf{E}}_{a \in_R A_1} \left[\sum_{x' \in X} |p_x(x') - f_p(x')| \right] = \text{dist}(p_x, f_p)\ . \tag{14}$$

Combining Equations (12), (13) and (14) together gives

$$\left| \mathop{\mathbf{E}}_{a \in_R A_1} [(M \cdot p_x)_a] \right| \ \geq\ 2^{-r} + 2^{-\xi+1} - \text{dist}(p_x, f_p)\ >\ 2^{-r}$$

Thus, we have two distributions p_a and p_x with min-entropy at least $\log(|A|) - k$ and $\log(|X|) - \ell$ respectively contradicting the fact that M is a $(k, \ell \sim r)$-L_∞-Extractor. Hence no such f exists and M is a $(k-1, \frac{\ell-\xi-1}{2})$-$L_2$-Extractor with error $2^{-r} + 2^{-\xi+1}$. □

5.1.4 Transpose

Lemma 5.5. *If a matrix $M : A \times X \to \{-1, 1\}$ is a (k, ℓ)-L_2-Extractor with error 2^{-r}, then the transposed matrix M^t is an $(\Omega(\ell), \Omega(k))$-$L_2$-Extractor with error $2^{-\Omega(\min\{r,k\})}$.*

Proof. As M is a (k, ℓ)-L_2-Extractor with error 2^{-r}, using Lemma 5.3, M is also a $(\Omega(k), \Omega(\ell) \sim (\Omega(\min\{r, k\})))$-$L_\infty$-Extractor. The definition of L_∞-Extractor is symmetric in its rows and columns and hence, M^t is also a $(\Omega(\ell), \Omega(k) \sim (\Omega(\min\{r, k\})))$-$L_\infty$-Extractor. Now, using Lemma 5.4 on M^t, we get that M^t is also a $(\Omega(\ell), \Omega(k))$-$L_2$-Extractor with error $2^{-\Omega(\min\{r,k\})}$. □

5.1.5 Lower Bounds for Almost Orthogonal Vectors

In this section, we show that a matrix $M : A \times X \to \{-1, 1\}$ whose rows are almost orthogonal is a good L_2-extractor. A similar technique was used in many previous works (see for example [GS71, CG88, A95, R05]). Motivated by the applications (e.g., learning sparse parities and learning from low-degree equations) in which some pairs of rows are not almost orthogonal, we relax this notion and only require that almost all pairs of rows are almost orthogonal. We formalize this in the definition of (ϵ, δ)-almost orthogonal vectors.

Definition 5.6. **(ϵ, δ)-almost orthogonal vectors:** *Vectors $v_1, \ldots, v_m \in \{-1, 1\}^X$ are* (ϵ, δ)-almost orthogonal *if for any $i \in [m]$ there are at most $\delta \cdot m$ indices $j \in [m]$ with $|\langle v_i, v_j\rangle| > \epsilon$.*

Definition 5.6 generalizes the definition of an (ϵ, δ)-biased set from [KRT16].

Definition 5.7. (ϵ, δ)-biased set ([KRT16]): *A set $T \subseteq \{0,1\}^n$ is (ϵ, δ)-*biased *if there are at most $\delta \cdot 2^n$ elements $a \in \{0,1\}^n$ with $|\mathbf{E}_{x \in_R T}[(-1)^{a \cdot x}]| > \epsilon$, (where $a \cdot x$ denotes the inner product of a and x, modulo 2).*

Definition 5.7 is a special case of Definition 5.6, where the vectors corresponding to a set $T \subseteq \{0,1\}^n$ are defined as follows. With every $a \in \{0,1\}^n$, we associate the vector v_a of length $|T|$, whose x-th entry equals $(-1)^{a \cdot x}$ for any $x \in T$. Indeed, T is (ϵ, δ)-biased iff the vectors $\{v_a : a \in \{0,1\}^n\}$ are (ϵ, δ)-almost orthogonal.

Lemma 5.8 (Generalized Johnson's Bound)**.** *Let $M \in \{-1,1\}^{A \times X}$ be a matrix. Assume that $\{M_a\}_{a \in A}$ are (ϵ, δ)-almost orthogonal vectors. Then, for any $\gamma > \sqrt{\epsilon}$ and any non-negative function $f : X \to \mathbb{R}^+$, we have at most $(\frac{\delta}{\gamma^2 - \epsilon}) \cdot |A|$ rows $a \in A$ with*

$$|\langle M_a, f \rangle| \geq \gamma \cdot \|f\|_2.$$

In particular, fixing $\gamma = \sqrt{\epsilon + \delta^{1/2}}$, we have that M is a (k, ℓ)-L_2-extractor with error 2^{-r}, for $k = \frac{1}{2}\log(1/\delta)$, and $\ell = r = \Omega\big(\min\{\log(1/\epsilon), \log(1/\delta)\}\big)$.

Proof. Fix $\gamma > \sqrt{\epsilon}$. Let I_+ (respectively, I_-) be the rows in A with high correlation (respectively, anti-correlation) with f. More precisely:

$$\begin{aligned} I_+ &:= \{i \in A : \ \langle M_i, f \rangle > \gamma \cdot \|f\|_2\}, \\ I_- &:= \{i \in A : \ -\langle M_i, f \rangle > \gamma \cdot \|f\|_2\}. \end{aligned}$$

Let $I = I_+ \cup I_-$. Define $z = \sum_{i \in I_+} M_i - \sum_{i \in I_-} M_i$. We consider the inner product of f and z. We have

$$\begin{aligned} (|I| \cdot \gamma \cdot \|f\|_2)^2 < \langle f, z \rangle^2 &= \left(\mathop{\mathbf{E}}_{x \in_R X} \left[f(x) \cdot \Big(\sum_{i \in I_+} M_{i,x} - \sum_{i \in I_-} M_{i,x} \Big) \right] \right)^2 \\ &\leq \mathop{\mathbf{E}}_{x \in_R X} \left[f(x)^2 \right] \cdot \mathop{\mathbf{E}}_{x \in_R X} \left[\Big(\sum_{i \in I_+} M_{i,x} - \sum_{i \in I_-} M_{i,x} \Big)^2 \right] \\ &\qquad \text{(Cauchy-Schwarz)} \\ &\leq \|f\|_2^2 \cdot \sum_{i \in I} \sum_{i' \in I} |\langle M_i, M_{i'} \rangle|. \end{aligned}$$

For any fixed $i \in I$, we break the inner-sum $\sum_{i' \in I} |\langle M_i, M_{i'} \rangle|$ according to whether or not $|\langle M_i, M_{i'} \rangle| > \epsilon$. By the assumption on M, there are at most $\delta \cdot |A|$ rows i' for which the inner-product is larger than ϵ. For these rows, the inner-product is at most 1. Thus, we get

$$(|I| \cdot \gamma \cdot \|f\|_2)^2 < \|f\|_2^2 \cdot \sum_{i \in I} \sum_{i' \in I} |\langle M_i, M_{i'} \rangle| \leq \|f\|_2^2 \cdot |I| \cdot (|A| \cdot \delta + \epsilon \cdot |I|).$$

That is,

$$|I| \cdot \gamma^2 < |A| \cdot \delta + \epsilon \cdot |I|.$$

Rearranging gives

$$|I| < \left(\frac{\delta}{\gamma^2 - \epsilon}\right) \cdot |A|,$$

which completes the first part of the proof.

We turn to the in particular part. Assume that $\frac{\|f\|_2}{\|f\|_1} \leq 2^\ell$. Thus, we proved that there are at most $\left(\frac{\delta}{\gamma^2-\epsilon}\right) \cdot |A|$ rows $a \in A$, such that,

$$|\langle M_a, f\rangle| \geq \gamma \cdot 2^\ell \cdot \|f\|_1.$$

Fixing $\gamma = \sqrt{\epsilon + \delta^{1/2}}$, $k = \log(1/\delta^{1/2})$, and $\ell = r = \frac{1}{2}\log(1/\gamma)$, we get that M is a (k,ℓ)-L_2-extractor with error 2^{-r} (Definition 2.1). Finally, note that $\ell = r = \Omega\big(\min\{\log(1/\delta), \log(1/\epsilon)\}\big)$, which completes the proof. □

5.2 Learning Sparse Parities

As an application of Lemma 5.8 and Theorem 1, we reprove the main result in [KRT16].

Lemma 5.9. *Let $T \subseteq \{0,1\}^n$ be an (ϵ,δ)-biased set, with $\epsilon \geq \delta$. Define the matrix $M : \{0,1\}^n \times T \to \{-1,1\}$ by $M(a,x) = (-1)^{a\cdot x}$. Then, the learning task associated with M ("parity learning over T") requires either at least $\Omega(\log(1/\epsilon) \cdot \log(1/\delta))$ memory bits or at least $\mathrm{poly}(1/\epsilon)$ samples.*

Proof. The rows $\{M_a\}_{a\in\{0,1\}^n}$ are (ϵ,δ)-almost orthogonal vectors. Thus, by Lemma 5.8, we get that M is a (k,ℓ)-L_2-extractor with error 2^{-r}, for $k = \Omega(\log(1/\delta))$ and $r = \ell = \Omega(\log(1/\epsilon))$ (assuming $\epsilon \geq \delta$). By Theorem 1, we get the required memory-samples lower bound. □

Lemma 5.10 ([KRT16])**.** *There exists a (sufficiently small) constant $c > 0$ such that the following holds. Let $T_\ell = \{x \in \{0,1\}^n : \sum_i x_i = \ell\}$. For any $\epsilon > (8\ell/n)^{\ell/2}$, T_ℓ is an (ϵ,δ)-biased set for $\delta = 2 \cdot e^{-\epsilon^{2/\ell}\cdot n/8}$. In particular, T_ℓ is an (ϵ,δ)-biased set for*

1. *$\epsilon = 2^{-c\ell}$, $\delta = 2^{-cn}$, assuming $\ell \leq cn$.*
2. *$\epsilon = \ell^{-c\ell}$, $\delta = 2^{-cn/\ell^{0.01}}$, assuming $\ell \leq n^{0.9}$.*

Let $c > 0$ be the constant mentioned in Lemma 5.10. The following lemma complements Lemma 5.10 to the range of parameters $cn \leq \ell \leq n/2$. It shows that T_ℓ is $(2^{-\Omega(n)}, 2^{-\Omega(n)})$-biased in this case. The proof is a simple application of Parseval's identity (see [KRT16]).

Lemma 5.11 ([KRT16, Lemma 4.1])**.** *Let $T \subseteq \{0,1\}^n$ be any set. Then, T is an (ϵ,δ)-biased set for $\delta = \frac{1}{|T|\cdot\epsilon^2}$. In particular, T is $(|T|^{-1/3}, |T|^{-1/3})$-biased.*

We get the following as an immediate corollary.

Corollary 4. *Let $T_\ell = \{x \in \{0,1\}^n : \sum_i x_i = \ell\}$.*

1. *Assuming $\ell \leq n/2$, parity learning over T_ℓ requires either at least $\Omega(n \cdot \ell)$ memory bits or at least $2^{\Omega(\ell)}$ samples.*
2. *Assuming $\ell \leq n^{0.9}$, parity learning over T_ℓ requires either at least $\Omega(n \cdot \ell^{0.99})$ memory bits or at least $\ell^{\Omega(\ell)}$ samples.*

5.3 Learning from Sparse Linear Equations

Lemma 5.5 and the proof of Lemma 5.9 gives the following immediate corollary.

Lemma 5.12. *Let $T \subseteq \{0,1\}^n$ be an (ϵ, δ)-biased set, with $\epsilon \geq \delta$. Then, the matrix $M : T \times \{0,1\}^n \to \{-1,1\}$, defined by $M(a,x) = (-1)^{a \cdot x}$ is a (k,ℓ)-L_2-extractor with error 2^{-r}, for $\ell = \Omega(\log(1/\delta))$ and $k = r = \Omega(\log(1/\epsilon))$.*

Thus, the learning task associated with M ("learning from equations in T") requires either at least $\Omega(\log(1/\epsilon) \cdot \log(1/\delta))$ memory bits or at least $\mathrm{poly}(1/\epsilon)$ samples.

We get the following as an immediate corollary of Lemmas 5.10, 5.11 and 5.12.

Corollary 5. *Let $T_\ell = \{x \in \{0,1\}^n : \sum_i x_i = \ell\}$.*

1. *Assuming $\ell \leq n/2$, learning from equations in T_ℓ requires either at least $\Omega(n \cdot \ell)$ memory bits or at least $2^{\Omega(\ell)}$ samples.*
2. *Assuming $\ell \leq n^{0.9}$, learning from equations in T_ℓ requires either at least $\Omega(n \cdot \ell^{0.99})$ memory bits or at least $\ell^{\Omega(\ell)}$ samples.*

5.4 Learning from Low Degree Equations

In the following, we consider multilinear polynomials in $\mathbb{F}_2[x_1, \ldots, x_n]$ of degree at most d. We denote by P_d the linear space of all such polynomials. We denote the bias of a polynomial $p \in \mathbb{F}_2[x_1, \ldots, x_n]$ by

$$\mathrm{bias}(p) := \mathop{\mathbf{E}}_{x \in \mathbb{F}_2^n}[(-1)^{p(x)}].$$

We rely on the following result of Ben-Eliezer, Hod and Lovett [BEHL12], showing that random low-degree polynomials have very small bias with very high probability.

Lemma 5.13 ([BEHL12, Lemma 2])**.** *Let $d \leq 0.99 \cdot n$. Then,*

$$\Pr_{p \in_R P_d}[|\mathrm{bias}(p)| > 2^{-c_1 \cdot n/d}] \leq 2^{-c_2 \cdot \binom{n}{\leq d}}$$

where $0 < c_1, c_2 < 1$ are absolute constants.

Corollary 6. *Let $d, n \in \mathbb{N}$, with $d \leq 0.99 \cdot n$. Let $M : P_d \times \mathbb{F}_2^n \to \{-1,1\}$ be the matrix defined by $M(p,x) = (-1)^{p(x)}$ for any $p \in P_d$ and $x \in \mathbb{F}_2^n$. Then, the vectors $\{M_p : p \in P_d\}$ are (ϵ, δ)-almost orthogonal, for $\epsilon = 2^{-c_1 n/d}$ and $\delta = 2^{-c_2 \binom{n}{\leq d}}$, (where $0 < c_1, c_2 < 1$ are absolute constants). In particular, M is a (k,ℓ)-L_2-extractor with error 2^{-r}, for $k = \Omega\left(\binom{n}{\leq d}\right)$ and $r = \ell = \Omega(n/d)$.*

Thus, the learning task associated with M ("learning from degree-d equations") requires either at least $\Omega\left(\binom{n}{\leq d} \cdot n/d\right) \geq \Omega((n/d)^{d+1})$ memory bits or at least $2^{\Omega(n/d)}$ samples.

Proof. We reinterpret [BEHL12, Lemma 2]. Since P_d is a linear subspace, for any fixed $p \in P_d$ and a uniformly random $q \in_R P_d$, we have that $p+q$ is a uniformly random polynomial in P_d. Thus, for any fixed $p \in P_d$, at most $2^{-c_2 \cdot \binom{n}{\leq d}}$ fraction of the polynomials $q \in P_d$ have

$$|\mathrm{bias}(p+q)| \geq 2^{-c_1 \cdot n/d}.$$

In other words, we get that $\{M_p : p \in P_d\}$ are (ϵ, δ)-almost orthogonal vectors for $\epsilon = 2^{-c_1 \cdot n/d}$ and $\delta = 2^{-c_2 \cdot \binom{n}{\leq d}}$. We apply Lemma 5.8 to get the "in particular" part, noting that in our case $\Omega\big(\min\{\log(1/\epsilon), \log(1/\delta)\}\big) = \Omega(n/d)$. We apply Theorem 1 to get the "thus" part. □

5.5 Learning Low Degree Polynomials

Lemma 5.5 and Corollary 6 gives the following immediate corollary.

Corollary 7. *Let $d, n \in \mathbb{N}$, with $d \leq 0.99 \cdot n$. Let $M : \mathbb{F}_2^n \times P_d \to \{-1, 1\}$ be the matrix defined by $M(a, p) = (-1)^{p(a)}$ for any $p \in P_d$ and $a \in \mathbb{F}_2^n$. Then, M is a (k, ℓ)-L_2-extractor with error 2^{-r}, for $\ell = \Omega\big(\binom{n}{\leq d}\big)$ and $k = r = \Omega(n/d)$.*

Thus, the learning task associated with M ("learning degree-d polynomials") requires either at least $\Omega\left(\binom{n}{\leq d} \cdot n/d\right) \geq \Omega((n/d)^{d+1})$ memory bits or at least $2^{\Omega(n/d)}$ samples.

5.6 Relation to Statistical-Query-Dimension

Let $\mathcal{C}$ be a class of functions mapping A to $\{-1, 1\}$. The Statistical-Query-Dimension of $\mathcal{C}$, denoted $\mathrm{SQdim}(\mathcal{C})$, is defined to be the maximal m such that there exist functions $f_1, \ldots, f_m \in \mathcal{C}$ with $|\langle f_i, f_j \rangle| \leq 1/m$ for all $i \neq j$ [K98, BFJKMR94]. As a corollary of Lemma 5.5 and Lemma 5.8, we get the following.

Corollary 8. *Let $\mathcal{C}$ be a class of functions mapping A to $\{-1, 1\}$. Let $\mathrm{SQdim}(\mathcal{C}) = m$. Let $f_1, \ldots, f_m \in \mathcal{C}$ with $|\langle f_i, f_j \rangle| \leq 1/m$ for any $i \neq j$. Define the matrix $M : A \times [m] \to \{-1, 1\}$ whose columns are the vectors $f_1, \ldots, f_m$. Then, M is a (k, ℓ)-L_2-extractor with error 2^{-r} for $k = \ell = r = \Omega(\log m)$.*

Thus, the learning task associated with M requires either at least $\Omega(\log^2 m)$ memory bits or at least $m^{\Omega(1)}$ samples.

Proof. Consider the rows of the matrix M^t. By our assumption, the rows of M^t are $(1/m, 1/m)$-almost orthogonal. Thus, by Lemma 5.8, M^t is a (k, ℓ)-L_2-extractor with error 2^{-r}, for $k = \ell = r = \Omega(\log m)$. By Lemma 5.5, M is a (k, ℓ)-L_2-extractor with error 2^{-r} for $k = \ell = r = \Omega(\log m)$. We apply Theorem 1 to get the "thus" part. □

In fact, we get the following (slight) generalization. Suppose that there are $m' \geq m$ functions $f_1, \ldots, f_{m'}$ mapping A to $\{-1, 1\}$ with $|\langle f_i, f_j \rangle| \leq 1/m$ for all $i \neq j$. Then, the learning task associated with the matrix whose columns are $f_1, \ldots, f_{m'}$ requires either at least $\Omega(\log(m) \cdot \log(m'))$ memory bits or at least $m^{\Omega(1)}$ samples.

5.7 Comparison with [R17]

Small Matrix Norm implies L_2-Extractor. This paper generalizes the result of [R17] that if a matrix $M : A \times X \to \{-1, 1\}$ is such that the largest singular value of M, $\sigma_{\max}(M)$, is at most $|A|^{\frac{1}{2}} |X|^{\frac{1}{2} - \varepsilon}$, then the learning problem represented by M requires either a memory of size at least $\Omega\left((\varepsilon n)^2\right)$ or at least $2^{\Omega(\varepsilon n)}$ samples, where $n = \log_2 |X|$. We use the following lemma:

Lemma 5.14. *If a matrix $M : A \times X \to \{-1, 1\}$ satisfies $\sigma_{\max}(M) \leq |A|^{\frac{1}{2}} \cdot |X|^{\frac{1}{2}-\varepsilon}$, then M is a (k, ℓ)-L_2-Extractor with error 2^{-r} for every $k, \ell, r > 0$ such that $k + 2\ell + 2r \leq 2\varepsilon n$ ($n = \log_2(|X|)$).*

Theorem 1 and Lemma 5.14 with $k = \varepsilon n, \ell = r = \frac{\varepsilon n}{4}$, imply the main result of [R17].

Proof. As $\sigma_{\max}(M) \leq |A|^{\frac{1}{2}}|X|^{\frac{1}{2}-\varepsilon}$, for a non-negative function $f : X \to \mathbb{R}$, $\|M \cdot f\|_2 \leq |X|^{1-\varepsilon} \cdot \|f\|_2$. In other words,

$$\left(\mathop{\mathbf{E}}_{a \in_R A}\left[|(M \cdot f)_a|^2\right]\right)^{1/2} \leq |X|^{1-\varepsilon} \cdot \|f\|_2$$

$$\implies \left(\mathop{\mathbf{E}}_{a \in_R A}\left[|\langle M_a, f\rangle|^2\right]\right)^{1/2} \leq |X|^{-\varepsilon} \cdot \|f\|_2$$

$$\implies \left(\mathop{\mathbf{E}}_{a \in_R A}\left[\left(\frac{|\langle M_a, f\rangle|}{\|f\|_1}\right)^2\right]\right)^{1/2} \leq 2^{-\varepsilon n} \cdot \frac{\|f\|_2}{\|f\|_1}$$

Now if $\frac{\|f\|_2}{\|f\|_1} \leq 2^{\ell}$ for some $\ell > 0$, then

$$\mathop{\mathbf{E}}_{a \in_R A}\left[\left(\frac{|\langle M_a, f\rangle|}{\|f\|_1}\right)^2\right] \leq 2^{-2\varepsilon n + 2\ell}.$$

Applying Markov's inequality, we get that there are at most $2^{-2\varepsilon n+2\ell+2r} \cdot |A|$ rows $a \in A$ with $\frac{|\langle M_a, f\rangle|}{\|f\|_1} \geq 2^{-r}$. □

5.8 Comparison with [MM17b]

We will now show that our result subsumes the one of [MM17b]. Moshkovitz and Moshkovitz [MM17b] consider matrices $M : A \times X \to \{-1, 1\}$, and a parameter d, with the property that for any $A' \subseteq A$ and $X' \subseteq X$ the bias of the submatrix $M_{A' \times X'}$ is at most $\frac{d}{\sqrt{|A'| \cdot |X'|}}$. They define $m = \frac{|A| \cdot |X|}{d^2}$ and prove that any learning algorithm for the corresponding learning problem requires either a memory of size $\Omega((\log m)^2)$ or $m^{\Omega(1)}$ samples. We note that this is essentially the same result as the one proved in [R17], and since it is always true that $d^2 \geq \max\{|X|, |A|\}$, the bound obtained on the memory is at most $\Omega\left(\min\left\{(\log |X|)^2, (\log |A|)^2\right\}\right)$.

Note that if M satisfies that property (required by [MM17b]), then, in particular, any submatrix $A' \times X'$ of M of at least $m^{-1/4} \cdot |A|$ rows and at least $m^{-1/4} \cdot |X|$ columns, has a bias of at most

$$\frac{d}{\sqrt{|A'| \cdot |X'|}} = \frac{d}{\sqrt{|A| \cdot |X|}} \cdot \frac{\sqrt{|A| \cdot |X|}}{\sqrt{|A'| \cdot |X'|}} \leq m^{-1/2} \cdot m^{1/4} = m^{-1/4}.$$

Thus, we can apply Corollary 3, with $k, \ell, r = \frac{1}{4}\log(m)$ to obtain the same result.

References

[A95] Noga Alon: Tools from Higher Algebra. In Handbook of Combinatorics, R.L.Graham, M.Grotschel and L.Lovasz, eds, North Holland (1995), Chapter 32: 1749-1783 24

[BEHL12] Ido Ben-Eliezer, Rani Hod, Shachar Lovett: Random low-degree polynomials are hard to approximate. Computational Complexity, 21(1): 63–81 (2012) 27

[BFJKMR94] Avrim Blum, Merrick L. Furst, Jeffrey C. Jackson, Michael J. Kearns, Yishay Mansour, Steven Rudich: Weakly learning DNF and characterizing statistical query learning using Fourier analysis. STOC 1994: 253-262 4, 28

[BL06] Yonatan Bilu, Nathan Linial: Lifts, Discrepancy and Nearly Optimal Spectral Gap. Combinatorica 26(5): 495-519 (2006) 2

[BOGY17] Paul Beame, Shayan Oveis Gharan, Xin Yang: Time-Space Tradeoffs for Learning from Small Test Spaces: Learning Low Degree Polynomial Functions. Manuscript (2017) 5

[CG88] Benny Chor, Oded Goldreich: Unbiased Bits from Sources of Weak Randomness and Probabilistic Communication Complexity. SIAM J. Comput. 17(2): 230-261 (1988) 2, 3, 22, 24

[GS71] Ronald Graham, Joel Spencer: A Constructive Solution to a Tournament Problem. Canad. Math. Bull. 14: 45-48 (1971) 24

[K98] Michael J. Kearns: Efficient Noise-Tolerant Learning from Statistical Queries. J. ACM 45(6): 983-1006 (1998) 4, 28

[KR13] Gillat Kol, Ran Raz: Interactive channel capacity. STOC 2013: 715-724 8

[KRT16] Gillat Kol, Ran Raz, Avishay Tal: Time-Space Hardness of Learning Sparse Parities. STOC 2017: 1067-1080 1, 2, 3, 4, 8, 25, 26

[MM17a] Dana Moshkovitz, Michal Moshkovitz: Mixing Implies Lower Bounds for Space Bounded Learning. Proceedings of the 2017 Conference on Learning Theory, PMLR 65:1516-1566, 2017. Also in: Electronic Colloquium on Computational Complexity (ECCC) 24: 17 (2017) 2, 4

[MM17b] Dana Moshkovitz, Michal Moshkovitz: Mixing Implies Strong Lower Bounds for Space Bounded Learning. Electronic Colloquium on Computational Complexity (ECCC) 24: 116 (2017) 1, 2, 4, 29

[MT17] Michal Moshkovitz, Naftali Tishby: Mixing Complexity and its Applications to Neural Networks. CoRR abs/1703.00729 (2017) 4

[R05] Ran Raz: Extractors with weak random seeds. STOC 2005: 11-20 24

[R16] Ran Raz: Fast Learning Requires Good Memory: A Time-Space Lower Bound for Parity Learning. FOCS 2016: 266-275 1, 2, 3, 4

[R17] Ran Raz: A Time-Space Lower Bound for a Large Class of Learning Problems. FOCS 2017 (to appear). Also in: Electronic Colloquium on Computational Complexity (ECCC) 24: 20 (2017) 1, 2, 3, 4, 5, 7, 8, 28, 29

[S14] Ohad Shamir: Fundamental Limits of Online and Distributed Algorithms for Statistical Learning and Estimation. NIPS 2014: 163-171 1, 4

[SV84] Miklos Santha, Umesh V. Vazirani: Generating Quasi-Random Sequences from Slightly-Random Sources. FOCS 1984: 434-440 2

[SVW16] Jacob Steinhardt, Gregory Valiant, Stefan Wager: Memory, Communication, and Statistical Queries. COLT 2016: 1490-1516 1, 4

[VV16] Gregory Valiant, Paul Valiant: Information Theoretically Secure Databases. Electronic Colloquium on Computational Complexity (ECCC) 23: 78 (2016) 1, 4

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il