Entropy Samplers and Strong Generic Lower Bounds For Space Bounded Learning*†

Dana Moshkovitz‡[1] **and Michal Moshkovitz**§[2]

1 **Department of Computer Science, UT Austin, USA**
danama@cs.utexas.edu

2 **The Edmond and Lily Safra Center for Brain Sciences, Hebrew University, Israel**
michal.moshkovitz@mail.huji.ac.il

Abstract

With any hypothesis class one can associate a bipartite graph whose vertices are the hypotheses $\mathcal{H}$ on one side and all possible labeled examples $\mathcal{X}$ on the other side, and an hypothesis is connected to all the labeled examples that are consistent with it. We call this graph the *hypotheses graph*. We prove that any hypothesis class whose hypotheses graph is mixing cannot be learned using less than $\Omega(\log^2|\mathcal{H}|)$ memory bits unless the learner uses at least a large number $|\mathcal{H}|^{\Omega(1)}$ labeled examples. Our work builds on a combinatorial framework that we suggested in a previous work for proving lower bounds on space bounded learning. The strong lower bound is obtained by defining a new notion of pseudorandomness, the entropy sampler. Raz obtained a similar result using different ideas.

1998 ACM Subject Classification I.2.6 Learning, F.1.3 Complexity Measures and Classes

Keywords and phrases learning, space bound, mixing, certainty, entropy sampler

Digital Object Identifier 10.4230/LIPIcs.ITCS.2018.28

1 Introduction

Let $\mathcal{H}$ be a family of Boolean hypotheses. One can learn an hypothesis from $\mathcal{H}$ after seeing $O(\log|\mathcal{H}|)$ random labeled examples. Intuitively, this is true since a typical labeled example cuts the number of possible hypotheses by a factor of two. However, learning with so few examples requires enough memory to store $\Theta(\log|\mathcal{H}|)$ examples in memory. If $\mathcal{X}$ is the family of possible labeled examples, then such a learner uses $\Theta(\log|\mathcal{X}|\cdot\log|\mathcal{H}|)$ memory bits. It is also possible to learn $\mathcal{H}$ using many fewer memory bits: enumerate the hypotheses one by one, moving to the next hypothesis only after encountering a new labeled example that is inconsistent with the current hypothesis. Such a brute force learner uses only $\log|\mathcal{H}|$ memory bits but requires an extravagant number $\Theta(|\mathcal{H}|\log|\mathcal{H}|)$ of labeled examples. A natural question is whether one can learn with *both* $\ll\Theta(\log|\mathcal{X}|\cdot\log|\mathcal{H}|)$ memory bits and $\ll|\mathcal{H}|$ labeled examples.

* A preliminary version of this work appeared as ECCC TR17-116.
† A full version of the paper is available at https://eccc.weizmann.ac.il/report/2017/116/.
‡ This material is based upon work supported by the National Science Foundation under grants number 1218547 and 1648712.
§ This work is partially supported by the Gatsby Charitable Foundation, The Israel Science Foundation, and Intel ICRI-CI center. M.M. is grateful to the Harry and Sylvia Hoffman Leadership and Responsibility Program.

Perhaps surprisingly, Raz [8] showed that parities ($\mathcal{X} = \{0,1\}^n \times \{0,1\}$ and $\mathcal{H} = \{\oplus_{i \in I} x_i | I \subseteq \{1,\ldots,n\}\}$) cannot be learned unless the learner uses either $\Omega(\log|\mathcal{X}| \cdot \log|\mathcal{H}|) = \Omega(n^2)$ memory bits or $|\mathcal{H}|^{\Omega(1)} = 2^{\Omega(n)}$ labeled examples. Until recently, parities gave the only hypothesis classes known with strong lower bounds on space-bounded learning[1].

In this work we show that strong lower bounds hold for any hypothesis class that satisfies a natural combinatorial condition about the mixing of a graph associated with the class. This subsumes the result on parities and shows similar results for random classes and classes that correspond to error correcting codes [6]. Many other applications follow using the large body of research on combinatorial mixing (see, e.g., [2]). More details will appear in the full version of this paper.

An hypothesis class can be described by a bipartite graph whose vertices are the hypotheses $\mathcal{H}$ and the labeled examples $\mathcal{X}$, and whose edges connect every hypothesis $h \in \mathcal{H}$ to the labeled examples $(x,y) \in \mathcal{X}$ that are consistent with it, i.e., $h(x) = y$. We say that the hypothesis class is d-mixing if for any set of hypotheses $A \subseteq \mathcal{H}$ and any set $B \subseteq \mathcal{X}$ of labeled examples it holds that $||E(A,B)| - |A||B|/2| \leq \mathrm{d}\sqrt{|A||B|}$, where $E(A,B)$ is the set of edges between A and B in the hypotheses graph. For instance, for parities, $d = \Theta(\sqrt{|\mathcal{X}|})$ (see, e.g., [6]). We prove that mixing hypothesis classes admit strong lower bounds on space-bounded learning.

▶ **Theorem 1** (main theorem). *If the hypotheses graph is* d*-mixing,* $r := \frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}$ *and* $|\mathcal{H}|$ *are at least some constants, then any learning algorithm that outputs the underlying hypothesis with probability at least* $r^{-\Theta(1)}$ *must use at least* $\Omega(\log^2 r)$ *memory bits or* $r^{\Omega(1)}$ *labeled examples.*

A similar theorem holds if the learner only *approximately* learns the underlying hypothesis [6].

1.1 Related Work

In this work we rely on a combinatorial framework – henceforth referred to as the *low certainty framework* – that we introduced in a previous work for analyzing space-bounded learning [6]. In [6] the bound on the number of memory states was only $\approx |\mathcal{H}|^{1.25}$ (the bound on the number of labeled examples was the optimal $|\mathcal{H}|^{\Omega(1)}$). In between those two works (the current work and [6]) Raz [9] showed a lower bound of $\Omega(\log^2|\mathcal{H}|)$ on the number of memory bits (as in the current paper), relying on a spectral mixing condition instead of a combinatorial mixing condition. In a subsequent work, Garg, Raz and Tal [3], and, independently, Beame, Gharan and Yang [1], improved the lower bound to the optimal $|\mathcal{X}|^{\Omega(\log|\mathcal{H}|)}$.

1.2 Entropy Sampler

The key idea in the current work is a new notion of pseudorandomness, which we call the entropy sampler. Fix a probability distribution p over a space $\mathcal{M}$. For every element $m \in \mathcal{M}$ let its "entropy level" be $k_m = \log(1/p(m))$. The min-entropy of p is $\min_m k_m$. A sampler with multiplicative error is defined as follows:

[1] Kol, Raz and Tal [4] generalized Raz's work to parities on l variables out of n, showing that either $\Omega(n)$ memory bits or $2^{\Omega(l)}$ examples are needed, and for $l \leq n^{0.9}$, either $\Omega(nl^{0.99})$ memory bits or $l^{\Omega(l)}$ examples are needed. Note: (1) For small l there are learners with both $\ll |\mathcal{X}|^{\Omega(\log|\mathcal{H}|)} = n^{\Omega(nl)}$ memory states and $\ll |\mathcal{H}|^{\Omega(1)} = n^{\Omega(l)}$ examples [4]. (2) The work [4] implies lower bounds for classes that contain parities on l out of n variables. To get a result for interesting classes, like DNFs or decision trees, one can pick $l \approx \log n$, but then the lower bounds are weak.

▶ **Definition 2** (Sampler)**.** A bipartite graph $(\mathcal{M}, \mathcal{H}, E)$ is a sampler with multiplicative factor L, min-entropy k and error ε, if for every distribution p over $\mathcal{M}$ of min-entropy at least k, for every $H \subseteq \mathcal{H}$, $|H| \geq \varepsilon|\mathcal{H}|$,

$$\sum_{m \in \mathcal{M}} p(m) \cdot \frac{|E(m,H)|}{|E(m,\mathcal{H})|} \leq L \cdot \frac{|H|}{|\mathcal{H}|},$$

where $E(\cdot,\cdot)$ denotes the set of edges between given memories and hypotheses in the knowledge graph.

The parameters of the sampler L, ε, are typically related to the min-entropy k. The higher the min-entropy k is, the lower the sampling parameters are. However it's possible, e.g., that all elements are at high entropy levels, except for one, for the min-entropy to be low and for the sampling parameters to be high. An entropy samplerbenefits from elements of all entropy levels starting k; higher entropy levels contribute to better sampling. Formally:

▶ **Definition 3** (entropy sampler)**.** A bipartite graph $(\mathcal{M}, \mathcal{H}, E)$ is an entropy samplerwith multiplicative factor L, min-entropy k, error ε and benefit α if for every distribution p of min-entropy k, for every $H \subseteq \mathcal{H}$, $|H| \geq \varepsilon|\mathcal{H}|$,

$$\sum_{m \in \mathcal{M}} p(m) \cdot \frac{|E(m,H)|}{|E(m,\mathcal{H})|} \cdot 2^{\alpha \cdot k_m} \leq L \cdot \frac{|H|}{|\mathcal{H}|}.$$

Typically, pseudorandom objects can only be defined with respect to min-entropy, and therefore the notion of an entropy sampleris unusual and may have other applications.

1.3 Proof Outline

The proof of Theorem 1 builds on the low certainty framework of [6]. A key object in the framework is the *knowledge graph* of the algorithm at various time steps. The knowledge graph at a certain time step is a bipartite graph, where one side corresponds to memory states of the learning algorithm and the other side corresponds to the possible hypotheses in $\mathcal{H}$. There is an edge (m, h) between a memory state m and an hypothesis h for every sequence of labeled examples that is consistent with h and leads to m at the relevant time step. Note that when an hypothesis is picked uniformly at random, the neighborhood of a memory state corresponds to the probability distribution over the possible hypotheses conditioned on landing in the memory state at the relevant time step. In this respect, the knowledge graph captures exactly the knowledge of the algorithm about the underlying hypothesis at the time step.

In order to prove lower bounds, the work [6] shows that when the hypotheses graph is mixing and the space is sufficiently bounded, the knowledge graph remains "pseudorandom" throughout the execution of the algorithm. Unfortunately, the pseudorandomness property of [6] (analogous to an extractor property) no longer holds when we wish to rule out learners that can store whole labeled examples in memory. The main idea of the current work is to prove that the knowledge graph is instead an entropy sampler(or, rather, a version of Definition 3 that is suitable for the knowledge graph). We show this by induction on the time t in the execution of the algorithm. Every probability distribution over memories at time $t+1$ corresponds to a probability distribution over memories at time t. This distribution depends on the likelihood of transitions to the time $t+1$ memories. Roughly speaking, less likely transitions from time t to time $t+1$ may give a lot of information about the underlying hypothesis. The notion of an entropy samplerguarantees that even after taking the new information into account sampling still holds (the actual analysis is quite involved, partly because it takes irregularity into account).

2 Preliminaries

$\log(\cdot)$ always means $\log_2(\cdot)$. The following claims were proven in [6]:

▶ **Claim 4.** *Let p be a probability distribution over a set A with $\sum_{i\in A} p(i)^2 \leq r$. Then, for every $A' \subseteq A$ it holds that $\sum_{i\in A'} p(i) \leq \sqrt{|A'|r}$.*

▶ **Claim 5** (generalized law of total probability)**.** *For any events A, B and a partition of the sample space $C_1, \ldots, C_n$,*

$$\Pr(A|B) = \sum_i \Pr(A|B, C_i)\Pr(C_i|B).$$

▶ **Claim 6** (generalized Bayes' theorem)**.** *For any three events A, B, C,*

$$\Pr(A|B, C) = \Pr(B|A, C)\frac{\Pr(A|C)}{\Pr(B|C)}$$

▶ **Claim 7.** *Suppose $B_1, \ldots, B_n$ are some disjoint events. Then,*

$$\Pr(A|B_1 \cup \ldots \cup B_n) = \sum_{i=1}^{n} \Pr(A|B_i)\frac{\Pr(B_i)}{\Pr(B_1 \cup \ldots \cup B_n)}.$$

2.1 Mixing

For a bipartite graph (A, B, E), A are the left vertices and B are the right vertices. For sets $S \subseteq A, T \subseteq B$ let

$$E(S, T) = \{(a, b) \in E | a \in S, b \in T\}.$$

For $a \in A$ (and similarly for $b \in B$) the neighborhood of a is $\Gamma(a) = \{b \in B | (a, b) \in E\}$, and the degree of a is $d_a = |\Gamma(a)|$. If all d_a are equal, we say that the graph is d_a-left regular or just left regular. We similarly define right regularity.

▶ **Definition 8** (mixing)**.** We say that a bipartite graph (A, B, E) with average left degree $\bar{d}_A$ is d-mixing if for any $S \subseteq A, T \subseteq B$ it holds that

$$\left||E(S, T)| - \frac{|S||T|}{|B|/\bar{d}_A}\right| \leq \mathrm{d}\sqrt{|S||T|}$$

▶ **Definition 9** (sampler)**.** A bipartite graph (A, B, E) is an (ϵ, ϵ')-*sampler* if for every $T \subseteq B$ it holds that

$$\Pr_{a\in A}\left(\left|\frac{|\Gamma(a) \cap T|}{d_a} - \frac{|T|}{|B|}\right| > \epsilon\right) < \epsilon',$$

where a is sampled uniformly.

We say that a vertex $a \in A$ *samples* T correctly if $\left|\frac{|\Gamma(a)\cap T|}{d_a} - \frac{|T|}{|B|}\right| \leq \epsilon$. The sampler property implies that there are only a few vertices $S \subseteq A$ that do not sample T correctly.

▶ **Claim 10** (Mixing implies sampler)**.** *If a bipartite graph (A, B, E) is* d*-mixing and d_A-left regular then it is also an $(\epsilon, \frac{2\mathrm{d}^2|B|}{d_A^2\epsilon^2|A|})$-sampler for any $\epsilon > 0$. Specifically, if $d_A = |B|/2$ then the graph is an $(\epsilon, \frac{8\mathrm{d}^2}{|B||A|\epsilon^2})$-sampler for any $\epsilon > 0$.*

Proof. See Claim 13 in [6]. ◀

3 The Low Certainty Framework

In this section we will summarize the main components of the combinatorial framework presented in our earlier work [6].

3.1 Hypotheses Graph

The hypotheses graph associated with an hypothesis class $\mathcal{H}$ and labeled examples $\mathcal{X}$ is a bipartite graph whose vertices are hypotheses in $\mathcal{H}$ and labeled examples in $\mathcal{X}$, and whose edges connect every hypothesis $h \in \mathcal{H}$ to the labeled examples $(x, y) \in \mathcal{X}$ that are consistent with h, i.e., $h(x) = y$.

Let us explore a few examples of hypothesis classes with mixing property.

parity. The hypotheses in $PARITY(n)$ are all the vectors in $\{0,1\}^n$, and the labeled examples are $\{0,1\}^n \times \{0,1\}$ (i.e., $|\mathcal{H}| = 2^n$ and $|\mathcal{X}| = 2 \cdot 2^n$).

▶ **Lemma 11** (Lindsey's Lemma). *Let H be a $n \times n$ matrix whose entries are 1 or -1 and every two rows are orthogonal. Then, for any $S, T \subseteq [n]$,*

$$\left| \sum_{i \in S, j \in T} H_{i,j} \right| \leq \sqrt{|S||T|n}.$$

Lindsey's Lemma and Claim 11 from [6] imply that the hypotheses graph of $PARITY(n)$ is $O(\sqrt{|\mathcal{X}|})$-mixing.

random class. For each hypothesis h and an example x, we have $h(x) = 1$ with probability $1/2$. The hypotheses graph is a random bipartite graph. It is well known that this graph is mixing (see [5]).

We can rephrase Claim 10 for the hypotheses graph and get

▶ **Proposition 12.** *If a graph $(\mathcal{H}, \mathcal{X}, E)$ is d-mixing then it is also $(\epsilon, \frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|\epsilon^2})$-sampler for any $\epsilon > 0$.*

3.2 H-expander

The main notion of expansion we will use for the hypotheses graph is H-expander, as we define next (H stands for Hypotheses graph). This notion follows from mixing (Definition 8).

▶ **Definition 13** (H-expander). A left regular bipartite graph (A, B, E) with left degree d_A is an $(\alpha, \beta, \epsilon)$-*H-expander* if for every $T \subseteq B, S \subseteq A$, with $|S| \geq \alpha|A|, |T| \geq \beta|B|$ it holds that

$$\left| |E(S,T)| - \frac{|S||T|}{|B|/d_A} \right| \leq \epsilon|S||T|.$$

For example, the hypotheses graph $(\mathcal{H}, \mathcal{X}, E)$ is left regular with left degree $|\mathcal{X}|/2$, so in this case the denominator $|B|/d_A$ will be equal to 2.

Note the following simple observation that relates mixing and H-expander.

▶ **Proposition 14.** *If a graph $(\mathcal{H}, \mathcal{X}, E)$ is d-mixing then it is also $(\alpha, \beta, \frac{2\mathrm{d}}{\sqrt{\alpha|\mathcal{H}|\beta|\mathcal{X}|}})$ – H-expander, for any $\alpha, \beta \in (0,1)$.*

3.3 Knowledge Graph

▶ **Definition 15** (knowledge graph)**.** The *knowledge graph at time* t of a learning algorithm with memory states $\mathcal{M}$ for an hypothesis class $\mathcal{H}$ is a bipartite *multigraph* $G_t = (\mathcal{H}, \mathcal{M}, E_t)$ where an edge $(h, m) \in E_t$ corresponds to a series of t labeled examples $(x_1, y_1), \ldots, (x_t, y_t)$ with $h(x_i) = y_i$ for every $1 \leq i \leq t$ and the algorithm ends up in memory state m after receiving these t examples.

At each step we will remove a tiny fraction of the edges from the knowledge graph and we focus only on the memories M_t — denote this graph by G'_t. We can read off from this graph the probability $q_t(h, m)$ which indicates the probability that the algorithm reached memory m after t steps and all examples are labeled by h. The probability $q_t(h, m)$ is proportional to the number of edges $E'_t(m, h)$ between a memory m and an hypothesis h in the graph G'_t. We can also observe the conditional probability $q_t(m|h)$ which is the probability that the algorithm reached memory state m given that all the examples observed after t steps are consistent with hypothesis h. We can deduce the probability of a memory m: $q_t(m) = \sum_t q_t(m|h)q_t(h)$. We can also find the probability of a set of memories $M \subseteq \mathcal{M}$, $q_t(M) = \sum_{m \in M} q_t(m)$. If the algorithm, after t steps, is in memory state m, we can deduce the probability that the true hypothesis is h, $q_t(h|m) = \frac{q_t(m|h)q_t(h)}{q_t(m)}$.

3.4 Certainty

Throughout the analysis we will maintain a substantial set of memories $M_t \subseteq \mathcal{M}$ and a set of hypotheses $H_t \subseteq \mathcal{H}$. At time t we pick the underlying hypothesis uniformly from H_t and only consider memories in M_t. Initially, before any labeled example is received, $H_0 = \mathcal{H}$ and M_0 contains all the memories. At later times, H_t and M_t will exclude certain bad hypotheses and memories.

Certainty is a progress measure for the learning algorithm defined as follows:

▶ **Definition 16** (certainty)**.** The *certainty* of a memory m at time t is defined as

$$\sum_h q_t(h|m)^2.$$

The *average certainty* of a set of memories M at time t is defined as

$$cer^t(M) := \sum_{m \in M} q_t(m) \sum_h q_t(h|m)^2.$$

To simplify the notation we write $cer^t(m)$ when we mean $cer^t(\{m\}) = q_t(m) \sum_h q_t(h|m)^2$, i.e., the average certainty with the set $\{m\}$ of memories. We also define a weighted certainty using a weight vector w of length $|\mathcal{M}|$ and each coordinate in w is some value in $[0, 1]$ by

$$cer^t_w(M) = \sum_{m \in M} q_t(m) w_m \cdot q_t^2(h|m).$$

Note that if w is the all 1 vector then $cer^t_w(M) = cer^t(M)$.

At each time t we will focus only on memories that are not too certain, i.e., whose certainty is not much more then the average certainty. Using Markov's inequality we will prove that with high probability the algorithm only reaches these not-too-certain memories. Let us define this set more formally,

$$Bad^c_M = \left\{ m \in M \;\middle|\; \sum_h q_t^2(h|m) > c \cdot cer^t(M_t) \right\},$$

for some $c > 0$, that is of the order $|\mathcal{H}|^{\epsilon}$, for some small constant ϵ. Oftentimes, we will omit c when it is clear from the context. For all $t \geq 1$ we will make sure that M_t will not include Bad_M^c (and additional memories, as will be defined in later sections). The following claims are proved in [6].

▶ **Claim 17.** *For any $c > 0$ and time t, $q_t(Bad_M^c) \leq 1/c$*

There is an equivalent definition of certainty in terms of the certainty of the hypothesis, rather than the memory.

▶ **Claim 18.** *For each memory m, hypothesis h and time t*

$$q_t(m)q_t(h|m)^2 = q_t(h)q_t(h|m)q_t(m|h)$$

In particular we can prove

▶ **Claim 19.** *The average certainty is also equal to*

$$cer^t(M) = \sum_{h\in\mathcal{H}} q_t(h) \sum_{m\in M} q_t(h|m)q_t(m|h).$$

We can therefore define the certainty of an hypothesis h, when focusing on a set of memories M as

$$\sum_{m\in M} q_t(h|m)q_t(m|h)$$

Given the last claim in mind we define

$$Bad_H^c = \{h \in \mathcal{H} \mid \sum_{m\in M_t} q_t(m|h)q_t(h|m) > c \cdot cer^t(M_t)\}.$$

Oftentimes, we will omit c when it is clear from the context.

Define $H_1 = \mathcal{H}$ and for $t > 1$, $H_{t+1} = H_t \setminus Bad_H$. We will define the distribution over the hypotheses at time t by $q_t(h) = \frac{1}{|H_t|}$ if $h \in H_t$, else $q_t(h) = 0$. The next claim proves that H_t is large.

▶ **Claim 20.** *For any $c > 0$, $|H_{t+1}| \geq (1 - 1/c)|H_t|$.*

In the rest of the paper we will prove that the average certainty of M_t, even for a large $t \sim \log c$, will be at most $\frac{c}{|\mathcal{H}|}$, and later we choose $c \sim \log \frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}$.

▶ **Claim 21.** *Suppose that the learning algorithm ends after t steps, $|H_t| \geq 3$ and at most γ fraction of the edges were removed from the knowledge graph. Then, there is an hypothesis h such that the probability to correctly return it is at most*

$$3\sqrt{c \cdot cer^t(M_t)} + 3(1 - q_t(M_t)) + \gamma$$

3.5 Representative Labeled Examples

For each memory m at time t, a representative labeled example x is one with $q_t(x|m)$ equal roughly to $\frac{1}{|\mathcal{X}|}$. In particular, given m and the unlabeled example, the probability to guess the label is roughly $1/2$.

▶ **Definition 22.** Let m be a memory state at time t, and let $\epsilon^{rep} > 0$. We say that a labeled example x is ϵ^{rep}-representative at m if

$$\frac{1-\epsilon^{rep}}{|\mathcal{X}|} \leq q_{t+1}(x|m) \leq \frac{1+\epsilon^{rep}}{|\mathcal{X}|}$$

We denote the set of labeled examples that are not ϵ^{rep}-representative at m by $NRep(m, \epsilon^{rep})$.

In [6] a weaker notion of $NRep$ with some specific constant ϵ^{rep} was used.

▶ **Claim 23.** *Let m be a memory in the knowledge graph at time t with certainty bounded by r, i.e., $\sum_h q_t(h|m)^2 \leq r$, assuming the hypotheses graph is an $(\alpha, \beta, \epsilon) - H$-expander, $|NRep(m, 4\sqrt{\alpha|\mathcal{H}|r} + 4\epsilon)| \leq 2\beta$.*

We prove this claim in Section 3.5.1.

3.5.1 Auxiliary Claims

The next claim will imply an equivalent definition for $NRep$.

▶ **Claim 24.** *For any set of labeled examples $S \subseteq \mathcal{X}$ and a memory m it holds that*

$$q_{t+1}(S|m) = \sum_h \Pr(S|h) q_t(h|m).$$

Proof. Using Claim 5 we know that

$$\begin{aligned} q_{t+1}(S|m) &= \sum_h q_{t+1}(S|m,h) q_t(h|m) \\ &= \sum_h \Pr(S|h) q_t(h|m) \end{aligned}$$

◀

Using Claim 24, we know that the not-representative set $NRep(m, \epsilon^{rep})$ is also equal to

$$\left\{ x \in \mathcal{X} | \sum_{h \in \mathcal{H}} \Pr(x|h) q_t(h|m) < \frac{1-\epsilon^{rep}}{|\mathcal{X}|} \right\} \cup \left\{ x \in \mathcal{X} | \sum_{h \in \mathcal{H}} \Pr(x|h) q_t(h|m) > \frac{1+\epsilon^{rep}}{|\mathcal{X}|} \right\}.$$

We would like to prove that $NRep(m, \epsilon^{rep})$ is small for any memory with small certainty. Note that

$$q_t(h|m,x) \propto q_t(h|m) I_{(x,h)\in E},$$

where $I_{(x,h)\in E}$ means that x and h are connected in the hypotheses graph (this follows from Claim 6 with $A = \{h\}, B = \{x\}, C = \{m\}$ and $q_t(x|h,m) = q_t(x|h) = \frac{2}{|\mathcal{X}|} I_{(x,h)\in E}$). This probability distribution can be imagined as if it were constructed by taking the hypotheses graph and adding weight $q_t(h|m)$ to every hypothesis h. Keeping this observation in mind we need some new notation.

Suppose there is a weight w_i for each hypothesis in the hypotheses graph $(\mathcal{H}, \mathcal{X}, E)$. Then, define the weights between sets $S \subseteq \mathcal{H}$ and $T \subseteq \mathcal{X}$ by $w(S,T) := \sum_{s\in S, t\in T} w(s) I_{(s,t)\in E}$ and $w(S) := \sum_{s\in S} w(s)$. We would like to prove that even if there are weights on the hypotheses the hypotheses graph is still pseudo-random. More formally, we will use the following definition.

▶ **Definition 25.** A left regular bipartite graph (A, B, E) is (β, ϵ) – weighted-expander with weights $w_1, \ldots, w_{|A|}$, $\sum_i w_i = 1$, $\forall i, w_i \geq 0$, and left degree d_A if for every $S \subseteq A$ and $T \subseteq B, |T| \geq \beta|B|$ it holds that

$$\left| w(S,T) - \frac{w(S)}{|B|/d_A}|T| \right| \leq \epsilon|T|$$

The next claim proves that any H-expander is a also a weighted-expander assuming low ℓ_2^2 weights.

▶ **Claim 26.** *[see [6]] If the hypotheses graph $(\mathcal{H}, \mathcal{X}, E)$ is an $(\alpha, \beta, \epsilon)$ – H-expander and $\sum_{i=1}^{|\mathcal{H}|} w_i^2 \leq r$ then the hypotheses graph is a $(\beta, 2\epsilon + 2\sqrt{\alpha|\mathcal{H}|r})$ – weighted-expander with weights $w_1, \ldots, w_{|\mathcal{H}|}$.*

Next we will prove our main claim in this section.

Proof of Claim 23. Denote $\epsilon^* = 4\sqrt{\alpha|\mathcal{H}|r} + 4\epsilon$. Define $T_1 = \{x | \sum_{h\in\mathcal{H}} \Pr(x|h)q_t(h|m) < \frac{1-\epsilon^*}{|\mathcal{X}|}\}$ and define weights to hypotheses $w(h) = q_t(h|m)$. From the definition of T_1 we know that

$$\sum_{h\in\mathcal{H}, x\in T_1} \Pr(x|h)q_t(h|m) < \frac{|T_1|(1-\epsilon^*)}{|\mathcal{X}|}.$$

The left term is equal to

$$\sum_{h\in\mathcal{H}, x\in T_1} \frac{2}{|\mathcal{X}|} I_{(x,h)\in E} q_t(h|m) = w(\mathcal{H}, T_1)\frac{2}{|\mathcal{X}|}$$

Assume by a way of contradiction that $|T_1| \geq \beta|\mathcal{X}|$, then Claim 26 implies that

$$\begin{aligned} w(\mathcal{H}, T_1)\frac{2}{|\mathcal{X}|} &\geq \left(\frac{w(\mathcal{H})}{2}|T_1| - 2(\sqrt{\alpha|\mathcal{H}|r} + \epsilon)|T_1|\right)\frac{2}{|\mathcal{X}|} \\ &= \frac{|T_1|}{|\mathcal{X}|} - 2\sqrt{\alpha|\mathcal{H}|r}\frac{2|T_1|}{|\mathcal{X}|} - 2\epsilon\frac{2|T_1|}{|\mathcal{X}|}, \end{aligned}$$

where the equality follows from the fact that $w(\mathcal{H}) = 1$.

Thus

$$\frac{|T_1|(1-\epsilon^*)}{|\mathcal{X}|} > \frac{|T_1|}{|\mathcal{X}|} - 2\sqrt{\alpha|\mathcal{H}|r}\frac{2|T_1|}{|\mathcal{X}|} - 2\epsilon\frac{2|T_1|}{|\mathcal{X}|},$$

$$\Rightarrow 4\sqrt{\alpha|\mathcal{H}|r} + 4\epsilon > \epsilon^*.$$

But the latter contradicts the definition of ϵ^*. Hence we can deduce that $|T_1| < \beta|\mathcal{X}|$.

Similarly, define $T_2 = \{x | \sum_{h\in\mathcal{H}} \Pr(x|h)q_t(h|m) > \frac{1+\epsilon^*}{|\mathcal{X}|}\}$. Assume by a way of contradiction that $|T_2| \geq \beta|\mathcal{X}|$ then

$$\frac{(1+\epsilon^*)|T_2|}{|\mathcal{X}|} < \sum_{h\in\mathcal{H}} \Pr(T_2|h)q_t(h|m) \leq \frac{|T_2|}{|\mathcal{X}|} + 2\sqrt{\alpha|\mathcal{H}|r}\frac{2|T_2|}{|\mathcal{X}|} + 2\epsilon\frac{2|T_2|}{|\mathcal{X}|},$$

where the left inequality follows from the definition of T_2 and the right inequality follows from Claim 26. So again we conclude that $|T_2| < \beta|\mathcal{X}|$. ◀

3.6 Decomposition to Heavy and Many Steps

We show that the certainty does not increase much with a single step of the algorithm. To this end, we decompose almost all the transitions of the bounded space algorithm to two kinds: either a *heavy-sourced* or *many-sourced*. A heavy-sourced memory state at time $t+1$ is one to which the algorithm moves from a memory state at time t via any labeled example from a large family of labeled examples. A many-sourced memory state at time $t+1$ is one that has many possible time-t sources. We analyze each kind of transition separately using H-expansion and K-expansion. For more details see [6].

4 Knowledge Graph Remains Pseudorandom

In this section we define a pseudorandomness property for the knowledge graph and prove that the knowledge graph maintains it throughout the execution of the algorithm, provided that the certainty is low and the hypotheses graph is mixing. To complete the proof we use the pseudorandomness of the knowledge graph to deduce the main theorem by adapting the low certainty framework [6]. For details see [7].

▶ **Definition 27** (enlarging distribution)**.** We say that a distribution p over the memories is (β,γ)-enlarging with respect to a probability distribution q if for every memory m it holds that $p(m) \leq \frac{q(m)}{\beta}$ and if $p(m) > 0$ then $p(m) \geq \frac{q(m)}{\beta} \cdot \gamma$.

β and γ provide a certain measure of the entropy in p. As usual, it is useful to use a logarithmic scale to measure the entropy and our log scale will be with respect to a parameter γ_0 associated with the hypothesis class.

▶ **Definition 28** (entropy-level)**.** The (p,q,β,γ_0)-entropy-level of an element m is defined as

$$e_{\gamma_0}(m) = \log_{\gamma_0} \frac{p(m)\beta}{q(m)}.$$

In other words, if $p(m) = \frac{q_t(m)}{\beta}\gamma_0^i$, then $e_{\gamma_0}(m) = i$.

▶ **Definition 29** (entropy sampler)**.** We say that the knowledge graph G'_t is an $(\alpha,\beta,\ell,\gamma_0,k)-$ entropy sampler if for every $H \subseteq \mathcal{H}$ with $|H| \geq \alpha|\mathcal{H}|$ and every (β,γ_0^k)-enlarging distribution p it holds that

$$\sum_m \Pr(H|m)p(m)2^{e_{\gamma_0}(m)} \leq \ell \cdot \frac{|H|}{|\mathcal{H}|}$$

The usual definition of sampler with multiplicative error is

$$\sum_m \Pr(H|m)p(m) \leq \ell \cdot \frac{|H|}{|\mathcal{H}|}.$$

Our definition requires more and seeks to benefit from memory states whose probability is much lower than $q_t(m^t)/\beta$.

Denote by $S^{m^t,m^{t+1}} \subseteq \mathcal{X}$ the examples that cause the memory to change from m^t to m^{t+1}.

▶ **Claim 30.** *Let $t \geq 1$. Assume that the following conditions hold:*

1. *The hypotheses graph is* d*-mixing.*
2. *The graph G'_t is an $(\alpha',\beta',\ell,\gamma_0,k)-$ entropy sampler.*

3. *All the edges* (m^t, m^{t+1}) *with labeled example* x *in* G'_t *are representative, i.e.,* $q_{t+1}(x|m^t) \notin NRep(m^t, \epsilon^{rep})$.
4. *All memories have low certainty, i.e., for all* m^t *in* G'_t, $cer(m^t) \leq c \cdot cer^t(M_t)$ *and* $cer^t(M_t) \leq c/|\mathcal{H}|$.
5. $\beta' \geq \gamma_0^{k-1}$ *and* $\alpha' \geq 2^{k+2}\sqrt{\gamma_0} + 2^{k+2} \cdot c \cdot \sqrt{\frac{16}{\gamma_0^{11}} \cdot \frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}}$.
6. $\epsilon^{rep} \leq 1/2$, *and* $\gamma_0 \leq 1/16$.

Then, G'_{t+1} *is an* $\left(\alpha', \beta', \left(1 + 10\sqrt{\gamma_0} + 2\epsilon^{rep}\right)\ell, \gamma_0, k\right) -$ *entropy sampler*

Proof. We can define a distribution q_{t+1} over pairs $(m^t, S^{m^t,m^{t+1}})$ where m^t is a memory at time t and $S^{m^t,m^{t+1}} \subseteq \mathcal{X}$ is the set of labeled examples that lead from m^t to m^{t+1}, in the following way

$$q_{t+1}(m^t, S^{m^t,m^{t+1}}) := q_t(m^t) q_{t+1}(S^{m^t,m^{t+1}}|m^t).$$

Fix a β'-enlarging distribution p (with respect to q_{t+1}) over memories at time $t+1$ and denote its support by M_{t+1}. For each $m^{t+1} \in M_{t+1}$, denote $p(m^{t+1}) = \frac{q_{t+1}(m^{t+1})}{\beta'_{m^{t+1}}}$, for $\beta'_{m^{t+1}} \geq \beta'$. This induces the distribution $p(m^t, S^{m^t,m^{t+1}}) := \frac{q_t(m^t) q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}}$. Indeed,

$$p(m^{t+1}) = \frac{q_{t+1}(m^{t+1})}{\beta'_{m^{t+1}}} = \frac{\sum_{m^t} q_{t+1}(m^t, S^{m^t,m^{t+1}})}{\beta'_{m^{t+1}}} = \sum_{m^t} p(m^t, S^{m^t,m^{t+1}})$$

The probability that p induces on memories at time t is

$$p(m^t) := \sum_{m^{t+1}} p(m^t, S^{m^t,m^{t+1}}) = q_t(m^t) \sum_{m^{t+1}} \frac{q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}}.$$

Fix $H \subseteq \mathcal{H}$ with $|H| \geq \alpha'|\mathcal{H}|$. In order to prove the claim, we would like to bound the expression

$$\begin{aligned} &\sum_{m^{t+1} \in M_{t+1}} q_{t+1}(H|m^{t+1}) p(m^{t+1}) 2^{e_{\gamma_0}(m^{t+1})} \\ &\quad = \sum_{m^{t+1} \in M_{t+1}} q_{t+1}(H|m^{t+1}) p(m^{t+1}) 2^{\log_{\gamma_0} \frac{p(m^{t+1})\beta'}{q_{t+1}(m^{t+1})}} \end{aligned} \tag{1}$$

The proof consists of five steps:

Step 1: Rewrite Expression 1 in terms of memories at time t:

Since $p(m^{t+1}) = \frac{q_{t+1}(m^{t+1})}{\beta'_{m^{t+1}}}$, Expression (1) is equal to

$$\sum_{m^{t+1}\in M_{t+1}} q_{t+1}(H|m^{t+1})p(m^{t+1})2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}}$$

$$(\text{dfn. of } m^{t+1}) = \sum_{m^{t+1}\in M_{t+1}} q_{t+1}(H|\vee_{m^t}(m^t, S^{m^t,m^{t+1}}))p(m^{t+1})2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}}$$

$$(\text{Claim 7}) = \sum_{\substack{m^{t+1}\in M_{t+1}\\ m^t\in M_t}} q_{t+1}(H|m^t, S^{m^t,m^{t+1}})\frac{q_{t+1}(m^t, S^{m^t,m^{t+1}})}{q_{t+1}(m^{t+1})}p(m^{t+1})2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}}$$

$$(\text{dfn. of } p) = \sum_{\substack{m^{t+1}\in M\\ m^t\in M_t, h\in H}} q_{t+1}(h|m^t, S^{m^t,m^{t+1}})$$
$$\frac{q_{t+1}(m^t, S^{m^t,m^{t+1}})}{q_{t+1}(m^{t+1})}\frac{q_{t+1}(m^{t+1})}{\beta'_{m^{t+1}}}2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}}$$

$$= \sum_{\substack{m^{t+1}\in M\\ m^t\in M_t, h\in H}} q_{t+1}(h|m^t, S^{m^t,m^{t+1}})\frac{q_t(m^t)q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}}2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}}$$

$$(\text{Claim 6}) = \sum_{\substack{m^{t+1}\in M_{t+1}\\ m^t\in M_t, h\in H}} q_t(h|m^t)$$
$$\frac{q_{t+1}(S^{m^t,m^{t+1}}|m^t, h)}{q_{t+1}(S^{m^t,m^{t+1}}|m^t)}\frac{q_t(m^t)q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}}2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}}$$

$$(\text{dfn. of } q_{t+1}) = \sum_{m^t\in M_t, h\in H} q_t(h|m^t)q_t(m^t)\sum_{m^{t+1}\in M_{t+1}}\frac{\Pr(S^{m^t,m^{t+1}}|h)}{\beta'_{m^{t+1}}}2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}} \quad (2)$$

In the next steps we will prove that for most memories m^t and for most hypotheses h the term inside the outer sum in (2) is bounded, that is,

$$q_t(h|m^t)q_t(m^t)\sum_{m^{t+1}\in M_{t+1}}\frac{\Pr(S^{m^t,m^{t+1}}|h)}{\beta'_{m^{t+1}}}2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}} \lesssim q_t(h|m^t)p(m^t)2^{e_{\gamma_0}(m^t)} \quad (3)$$

Moreover, the effect of the other memories and hypothesis is negligible. Proving the latter will finish the proof since G'_t is a entropy sampler.

1. In step 2 we show that memories m_t with low $p(m^t)$ do not add much to Expression (2).
2. In step 3 we focus on a memory m_t whose $p(m^t)$ is now low. To show that Inequality (3) holds for most hypotheses h we first recall that since

$$p(m^t) = q_t(m^t)\sum_{m^{t+1}}\frac{q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}},$$

we need to prove that

$$\sum_{m^{t+1}}\frac{\Pr(S^{m^t,m^{t+1}}|h)}{\beta'_{m^{t+1}}}2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}} \lesssim \sum_{m^{t+1}}\frac{q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}}2^{e_{\gamma_0}(m^t)} \quad (4)$$

In step 3 we show that for most hypotheses h it holds that

$$\Pr(S^{m^t,m^{t+1}}|h) \sim \frac{|S^{m^t,m^{t+1}}|}{|\mathcal{X}|} \sim q_{t+1}(S^{m^t,m^{t+1}}|m^t).$$

3. In step 4 we show that the hypotheses that are not considered in the previous step do not add much to Expression (2).
4. In step 5 we would like to show that Inequality (4) holds. After step 3 and the definition of $e_{\gamma_0}(m)$ this is merely showing that

$$\sum_{m^{t+1}\in M_{t+1}} \frac{\Pr(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}} 2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}}$$

$$\lesssim \left(\sum_{m^{t+1}} \frac{q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}}\right) 2^{\log_{\gamma_0}\beta' \sum_{m^{t+1}} \frac{q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}}}$$

   This is proved in step 5 using Jensen's inequality.
5. In step 6 we sum everything up.

Step 2: getting rid of low p-weight memories at time t: In order to use the assumption in the claim regarding the entropy sampler property of G'_t, we need to make sure that for each memory m^t at time t, $p(m^t) = 0$ or $p(m^t) \geq \frac{q_t(m^t)}{\beta'/\gamma_0^k}$. Denote by Low the set of all memories m^t at time t with $0 < p(m^t) < \frac{q_t(m^t)}{\beta'/\gamma_0^k}$. Note that this set has low p-weight

$$p(Low) = \sum_{m^t\in Low} p(m^t) < \sum_{m^t\in Low} q_t(m^t)\frac{\gamma_0^k}{\beta'} \leq \frac{\gamma_0^k}{\beta'} \leq \gamma_0, \tag{5}$$

where the last inequality is true since $\beta' \geq \gamma_0^{k-1}$. Thus, by setting the probability of the memories in Low to 0, the remaining memories need to be multiplied by a factor of at most $1/(1-\gamma_0)$ (i.e., by a factor that is close to 1) so as to make it a distribution again. More formally, we divide the sum that we want to bound, Expression (2), into two sums depending on the membership in Low:

$$\sum_{m^t\in Low, h\in H} q_t(h|m^t)q_t(m^t) \sum_{m^{t+1}\in M_{t+1}} \frac{\Pr(S^{m^t,m^{t+1}}|h)}{\beta'_{m^{t+1}}} 2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}} +$$

$$+ \sum_{m^t\in M_t\setminus Low, h\in H} q_t(h|m^t)q_t(m^t) \sum_{m^{t+1}\in M_{t+1}} \frac{\Pr(S^{m^t,m^{t+1}}|h)}{\beta'_{m^{t+1}}} 2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}} \tag{6}$$

For $m^t \in Low$, the expression $2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}}$ is at most 2^k (since $\frac{q_{t+1}(m^{t+1})}{\beta'_{m^{t+1}}} = p(m^{t+1}) \geq$

$\frac{q_{t+1}(m^{t+1})\gamma_0^k}{\beta'}$ for any m^{t+1}). Thus, the first term in Expression (6) is at most

$$\begin{aligned}
&\sum_{m^t\in Low, h\in H} q_t(h|m^t)q_t(m^t) \sum_{m^{t+1}\in M_{t+1}} \frac{\Pr(S^{m^t,m^{t+1}}|h)}{\beta'_{m^{t+1}}}\cdot 2^k \\
(\text{see Claim 32}) &\leq \sum_{m^t\in Low, h\in H} q_t(h|m^t)q_t(m^t)\cdot \\
&\quad \sum_{m^{t+1}\in M_{t+1}} \frac{2(1+2\epsilon^{rep})q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}}\cdot 2^k \\
(\text{definition of } p(m^t)) &= \sum_{m^t\in Low} q_t(H|m^t)2^{k+1}(1+2\epsilon^{rep})p(m^t) \\
(q_t(H|m^t)\leq 1, \epsilon^{rep}\leq 1/2) &\leq 2^{k+2}\sum_{m^t\in Low} p(m^t) \\
(\text{see Inequality (5)}) &\leq 2^{k+2}\gamma_0 \qquad (7)
\end{aligned}$$

Denote $s = p(Low)$. The second term in Expression (6) is equal to

$$(1-s)\sum_{m^t\in M_t\setminus Low, h\in H} q_t(h|m^t)q_t(m^t) \sum_{m^{t+1}\in M_{t+1}} \frac{\Pr(S^{m^t,m^{t+1}}|h)}{(1-s)\beta'_{m^{t+1}}} 2^{\log_{\gamma_0}\frac{\beta'}{\frac{1-s}{1-s}\cdot\beta'_{m^{t+1}}}}$$

which is at most

$$\sum_{m^t\in M_t\setminus Low, h\in H} q_t(h|m^t)q_t(m^t) \sum_{m^{t+1}\in M_{t+1}} \frac{\Pr(S^{m^t,m^{t+1}}|h)}{(1-s)\beta'_{m^{t+1}}} 2^{\log_{\gamma_0}\frac{\beta'}{(1-s)\beta'_{m^{t+1}}}}\cdot 2^{\log_{\gamma_0} 1-s}$$

Using Claim 34, $\gamma_0 \leq 1/16$, and Inequality (5), it is at most

$$(1+\sqrt{\gamma_0})\sum_{m^t\in M_t\setminus Low, h\in H} q_t(h|m^t)q_t(m^t) \sum_{m^{t+1}\in M_{t+1}} \frac{\Pr(S^{m^t,m^{t+1}}|h)}{(1-s)\beta'_{m^{t+1}}} 2^{\log_{\gamma_0}\frac{\beta'}{(1-s)\beta'_{m^{t+1}}}} \qquad (8)$$

Step 3: $\Pr(S^{m^t,m^{t+1}}|h) \sim \frac{|S^{m^t,m^{t+1}}|}{|\mathcal{X}|} \sim q_{t+1}(S^{m^t,m^{t+1}}|m^t)$: Focus on a memory $m^t \notin Low$. In this step we will prove that for most hypotheses h the term $\Pr(S^{m^t,m^{t+1}}|h)$ can be replaced by $\Pr(S^{m^t,m^{t+1}}|m^t)$. We would like to rewrite the inner sum,

$$\sum_{m^{t+1}\in M_{t+1}} \frac{\Pr(S^{m^t,m^{t+1}}|h)}{\beta'_{m^{t+1}}} 2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}},$$

in Expression (2). For this purpose we first sort all the memories in $m^{t+1}\in M_{t+1}$ according to ascending order of $2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}}/\beta'_{m^{t+1}}$. Denote by β'_i the value $\beta'_{m^{t+1}}$ for m^{t+1} that is the i-th member in the sorted order. Then we get that the inner sum in Expression (2) is equal

to

$$\begin{aligned}\sum_{m^i\in M_{t+1}}\Pr(S^{m^t,m^i}|h)\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_i}}}{\beta'_i} &= \sum_{j\geq 1}\Pr(S^{m^t,m^j}|h)\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_1}}}{\beta'_1}+\\ &+\sum_{j\geq 2}\Pr(S^{m^t,m^j}|h)\left(\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_2}}}{\beta'_2}-\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_1}}}{\beta'_1}\right)+\\ &+\sum_{j\geq 3}\Pr(S^{m^t,m^j}|h)\left(\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_3}}}{\beta'_3}-\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_2}}}{\beta'_2}\right)+\dots\end{aligned}$$

Denote by $S^{m^t,\geq i}$ all the examples that lead from the memory m^t to any of the time-$(t+1)$ memories that are not the first $i-1$ memories. For convenience, define $1/\beta'_0 := 0$. Thus, it holds that

$$\sum_{m_i\in M_{t+1}}\Pr(S^{m^t,m_i}|h)\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_i}}}{\beta'_i}=\sum_{i\geq 1}\Pr(S^{m^t,\geq i}|h)\left(\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_i}}}{\beta'_i}-\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{i-1}}}}{\beta'_{i-1}}\right).$$

We divide this sum into two, using index $i_{(m^t)}$ which is the largest i such that $|S^{m^t,\geq i}|\geq\epsilon'|\mathcal{X}|$, for ϵ' to be determined later.

$$\begin{aligned}&\sum_{i=1}^{i_{(m^t)}}\Pr(S^{m^t,\geq i}|h)\left(\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_i}}}{\beta'_i}-\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{i-1}}}}{\beta'_{i-1}}\right)+\\ &\sum_{i=(i_{(m^t)})+1}^{|M_{t+1}|}\Pr(S^{m^t,\geq i}|h)\left(\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_i}}}{\beta'_i}-\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{i-1}}}}{\beta'_{i-1}}\right)\end{aligned}\tag{9}$$

Let us start with bounding the first term in Equation (9). From Claim 33, we know that except for a fraction of $\frac{1}{\epsilon^2}\cdot\frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}$ hypotheses $h\in\mathcal{H}$ for each $i\leq(1-\epsilon')|\mathcal{X}|$,

$$\Pr(S^{m^t,\geq i}|h)\leq\left(1+\epsilon'+\frac{4\epsilon}{(\epsilon')^2}\right)\frac{|S^{m^t,\geq i}|}{|\mathcal{X}|},\tag{10}$$

for $\epsilon>0$ to be determined later. From Claim 32 we know that the RHS is at most

$$\left(1+\epsilon'+\frac{4\epsilon}{(\epsilon')^2}\right)(1+2\epsilon^{rep})q_{t+1}(S^{m^t,\geq i}|m^t)$$

Denote the set of hypotheses that the bound in Inequality (10) does not apply to by $Err(m^t)$. We know that

$$\frac{|Err(m^t)|}{|\mathcal{H}|}\leq\frac{1}{\epsilon^2}\cdot\frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}\tag{11}$$

Let us now bound the second term in Expression (9). For each $i>i_{(m^t)}$ we use the simple bound given in Claim 32:

$$\Pr(S^{m^t,\geq i}|h)\leq 2(1+2\epsilon^{rep})q_{t+1}(S^{m^t,\geq i}|m^t).\tag{12}$$

We can now rewrite Expression (9) using Inequalities 10 and 12. Namely, for $m^t \notin Low$ and $h \notin Err(m^t)$ Expression (9) is at most

$$\left(1+\epsilon'+\frac{4\epsilon}{(\epsilon')^2}\right)(1+2\epsilon^{rep})\left[\sum_{i=1}^{i_{(m^t)}} q_{t+1}(S^{m^t,\geq i}|m^t)\left(\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_i}}}{\beta'_i}-\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{i-1}}}}{\beta'_{i-1}}\right)+\right.$$
$$\left.\sum_{i=(i_{(m^t)})+1}^{|M_{t+1}|} 2\cdot q_{t+1}(S^{m^t,\geq i}|m^t)\left(\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_i}}}{\beta'_i}-\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{i-1}}}}{\beta'_{i-1}}\right)\right]$$

Which is equal to

$$\left(1+\epsilon'+\frac{4\epsilon}{(\epsilon')^2}\right)(1+2\epsilon^{rep})\cdot$$
$$\left[\sum_{i=1}^{i_{(m^t)}} q_{t+1}(S^{m^t,m_i}|m^t)\frac{2^{\log_{\gamma_0}\frac{\beta'}{\beta'_i}}}{\beta'_i}+\sum_{i=(i_{(m^t)})+1}^{|M_{t+1}|} q_{t+1}(S^{m^t,m_i}|m^t)\frac{2\cdot 2^{\log_{\gamma_0}\frac{\beta'}{\beta'_i}}}{\beta'_i}\right] \tag{13}$$

Step 4: getting rid of "bad" hypotheses: We would like to bound the portion of Expression (2) that involves $h \in Err(m^t)$ for some m^t. Namely, we would like to bound

$$\sum_{m^t\in M_t, h\in Err(m^t)} q_t(h|m^t)q_t(m^t)\sum_{m^{t+1}\in M_{t+1}}\frac{\Pr(S^{m^t,m^{t+1}}|h)}{\beta'_{m^{t+1}}}2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}}. \tag{14}$$

For any m^{t+1}, from the definition of p we know that $\frac{q_{t+1}(m^{t+1})}{\beta'_{m^{t+1}}} = p(m^{t+1}) \geq \frac{q_{t+1}(m^{t+1})\gamma_0^k}{\beta'}$, hence $2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}} \leq 2^k$. Hence Expression (14) is at most

$$\sum_{m^t\in M_t, h\in Err(m^t)} q_t(h|m^t)q_t(m^t)\sum_{m^{t+1}\in M_{t+1}}\frac{\Pr(S^{m^t,m^{t+1}}|h)}{\beta'_{m^{t+1}}}2^k.$$

From Claim 32 we know that $\frac{\Pr(S^{m^t,m^{t+1}}|h)}{\beta'_{m^{t+1}}} \leq \frac{4q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}}$. Hence, Expression (14) is at most

$$\begin{aligned}\sum_{\substack{m^t\in M_t\\ h\in Err(m^t)}} q_t(h|m^t)q_t(m^t) & \sum_{m^{t+1}\in M_{t+1}}\frac{q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}}2^{k+2}\\ &= \sum_{\substack{m^t\in M_t\\ h\in Err(m^t)}} q_t(h|m^t)p(m^t)2^{k+2}\\ &\leq 2^{k+2}\sum_{m^t\in M_t} p(m^t)q_t(Err(m^t)|m^t).\end{aligned}$$

From Claim 4 and Inequality (11) we know that

$$q_t(Err(m^t)|m^t) \leq \sqrt{|Err(m^t)|c\cdot cer^t(M_t)} \leq c\cdot\sqrt{\frac{1}{\epsilon^2}\cdot\frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}},$$

where the second inequality follows from Inequality (11) and the assumption in the claim regarding the bound on $cer^t(M_t)$. To sum up this step, $Err(m^t)$ adds only a small additive error of $2^{k+2}\cdot c\cdot\sqrt{\frac{1}{\epsilon^2}\cdot\frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}}$ to Expression (2).

Step 5: towards using the entropy sampler property of G'_t: Recall that according to our plan at step 1 we want to prove now that for $m^t \notin Low, h \notin Err(m^t)$ it holds that

$$\sum_{m_j \in M_{t+1}} \frac{\Pr(S^{m^t,m_j}|m^t)}{\beta'_{m_j}} 2^{\log_{\gamma_0} \frac{\beta'}{\beta'_{m_j}}}$$

is at most

$$\left(\sum_{m_j} \frac{q_{t+1}(S^{m^t,m_j}|m^t)}{\beta'_{m_j}} \right) 2^{\log_{\gamma_0} \beta' \sum_{m_j} \frac{q_{t+1}(S^{m^t,m_j}|m^t)}{\beta'_{m_j}}} (1+\epsilon'_4)$$

for some small $\epsilon'_4 \in (0,1)$ to (implicitly) be determined in the next step. To this end we first prove, having in mind the expression in 13, that the following inequality holds

$$\frac{\sum_{i=1}^{i_{(m^t)}} \frac{q_{t+1}(S^{m^t,m_i}|m^t)}{\beta'_i} 2^{\log_{\gamma_0} \frac{\beta'}{\beta'_i}} + \sum_{i=(i_{(m^t)})+1}^{|M_{t+1}|} \frac{q_{t+1}(S^{m^t,m_i}|m^t)}{\beta'_i} 2^{\log_{\gamma_0} \frac{\beta' \cdot \gamma_0}{\beta'_i}}}{\sum_{m_j} \frac{q_{t+1}(S^{m^t,m_j}|m^t)}{\beta'_{m_j}}}$$
$$\leq 2^{\log_{\gamma_0} \beta' \sum_{m_j} \frac{q_{t+1}(S^{m^t,m_j}|m^t)}{\beta'_{m_j}}} (1+\epsilon_4) \tag{15}$$

for some small $\epsilon_4 \in (0,1)$ to be determined later.

Define the function $f(x) = 2^{\log_{\gamma_0} \frac{1}{x}}$ and the following distribution over memories at time $t+1$: $\bar{p}(m^i) \propto \frac{q_{t+1}(S^{m^t,m^i}|m^t)}{\beta'_{m^i}}$ and divide both sides by $2^{\log_{\gamma_0} \beta'}$ then Inequality (15) can be rewritten as

$$\sum_{m_i} \bar{p}(m_i) f\left(\beta'_i \cdot \left(\frac{1}{\gamma_0}\right)^{I_{i>i_{(m^t)}}} \right) \leq f\left(\left(\frac{1}{\gamma_0}\right)^{\log(1+\epsilon_4)} / \sum_{m_j} \frac{q_{t+1}(S^{m^t,m_j}|m^t)}{\beta'_{m_j}} \right),$$

where I is the indicator function. Use Jensen's inequality with the concave function f (see Claim 31) and get that the LHS is at most

$$f\left(\sum_{m_i} \frac{q_{t+1}(S^{m^t,m_i}|m^t)}{\sum_{m_j} \frac{q_{t+1}(S^{m^t,m_j}|m^t)}{\beta'_{m_j}}} \cdot \left(\frac{1}{\gamma_0}\right)^{I_{i>i_{(m^t)}}} \right)$$

Since f is monotonically increasing (see Claim 31), to prove Inequality (15) it is enough to show that

$$\sum_{m_i} q_{t+1}(S^{m^t,m_i}|m^t) \cdot \left(\frac{1}{\gamma_0}\right)^{I_{i>i_{(m^t)}}} \leq \left(\frac{1}{\gamma_0}\right)^{\log(1+\epsilon_4)}$$

Using the inequality $x/2 \leq \log(1+x)$ (which follows from Fact 35 and $\epsilon_4 < 1$) it is enough to prove that

$$\sum_{m_i} q_{t+1}(S^{m^t,m_i}|m^t) \cdot \left(\frac{1}{\gamma_0}\right)^{I_{i>i_{(m^t)}}} \leq \left(\frac{1}{\gamma_0}\right)^{\epsilon_4/2}. \tag{16}$$

Note that by separating the LHS into two and the definition of ϵ' we have that

$$\sum_{m_i} q_{t+1}(S^{m^t,m_i}|m^t)\cdot\left(\frac{1}{\gamma_0}\right)^{I_{i>i_{(m^t)}}} \leq 1+\sum_{i>i_{(m^t)}} q_{t+1}(S^{m^t,m_i}|m^t)\left(\frac{1}{\gamma_0}\right) \leq 1+\epsilon'\left(\frac{1}{\gamma_0}\right)$$

Thus, to show that Inequality (16) holds, it suffices to show that

$$1+\epsilon'\left(\frac{1}{\gamma_0}\right) \leq \left(\frac{1}{\gamma_0}\right)^{\epsilon_4/2}.$$

Which is true if and only if

$$\ln\left(1+\epsilon'\left(\frac{1}{\gamma_0}\right)\right) \leq \frac{\epsilon_4}{2}\ln\left(\frac{1}{\gamma_0}\right).$$

Using Fact 35 it is enough to show that

$$\epsilon'\left(\frac{1}{\gamma_0}\right) \leq \frac{\epsilon_4}{2}\ln\left(\frac{1}{\gamma_0}\right).$$

We choose $\epsilon_4 = 2\sqrt{\epsilon'}$. If $\sqrt{\epsilon'} \leq \gamma_0$ then the inequality will hold since $\gamma_0 \leq 1/16 < 1/e$.

Step 6: Summing up: Using Expressions (8), (13), (15) (recall that $\epsilon_4 = 2\sqrt{\epsilon'}$), the assumption is the claim regarding the entropy sampler of G'_t, Expression (7), and the conclusion of step 4 we have proven that Expression (1) is bounded by

$$(1+\sqrt{\gamma_0})\left(1+\epsilon'+\frac{4\epsilon}{(\epsilon')^2}\right)(1+2\epsilon^{rep})(1+2\sqrt{\epsilon'})\ell\cdot\frac{|H|}{|\mathcal{H}|}+2^{k+2}\gamma_0+2^{k+2}\cdot c\cdot\sqrt{\frac{1}{\epsilon^2}\cdot\frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}}$$

We choose $\epsilon' = \gamma_0^2$ (note that indeed $\sqrt{\epsilon'} \leq \gamma_0$) and $\epsilon = \gamma_0^5/4$. From the assumption in the claim we know that $\alpha'\sqrt{\gamma_0} \geq 2^{k+2}\gamma_0 + 2^{k+2}\cdot c\cdot\sqrt{\frac{16}{\gamma_0^{10}}\cdot\frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}}$. Hence, Expression (1) is at most

$$\left((1+\sqrt{\gamma_0})\left(1+\gamma_0^2+\gamma_0\right)(1+2\epsilon^{rep})(1+2\gamma_0)\ell+\sqrt{\gamma_0}\right)\cdot\frac{|H|}{|\mathcal{H}|} \leq (1+10\sqrt{\gamma_0}+2\epsilon^{rep})\,\ell\cdot\frac{|H|}{|\mathcal{H}|}$$

(in the RHS the constant 10 near $\sqrt{\gamma_0}$ was chosen arbitrarily) ◀

4.1 Auxiliary Claims

▶ **Claim 31.** *For any $\epsilon \leq 1/2$, the function $f(x) = 2^{\log_\epsilon \frac{1}{x}}$ for $x > 0$ is monotonically increasing and concave.*

In the next claim we lower bound $q_{t+1}(S|m^t)$ in terms of $\Pr(S|h)$ via the term $|S|/|\mathcal{X}|$.

▶ **Claim 32.** *Let $S \subseteq \mathcal{X}$. Let $h \in \mathcal{H}$.*

1. $\Pr(S|h) \leq \frac{2|S|}{|\mathcal{X}|}$
2. *Let m^t be a memory at time t. Assume $S \cap NRep(m^t, \epsilon^{rep}) = \emptyset$ and $\epsilon^{rep} \leq 1/2$. Then $\frac{|S|}{|\mathcal{X}|} \leq (1+2\epsilon^{rep})q_{t+1}(S|m^t)$.*

Proof. The first inequality follows from the fact that if $(x,h) \in E$ (i.e., hypothesis h and labeled example x are consistent) then $\Pr(x|h) = 2/|\mathcal{X}|$ and if $(x,h) \notin E$ then $\Pr(x|h) = 0$. To prove the second inequality, we use the definition of $NRep$ (see Definition 22) to deduce that

$$\frac{1-\epsilon^{rep}}{|\mathcal{X}|}|S| \leq q_{t+1}(S|m^t) \Rightarrow \frac{|S|}{|\mathcal{X}|} \leq \frac{1}{1-\epsilon^{rep}}q_{t+1}(S|m^t) \Rightarrow \frac{|S|}{|\mathcal{X}|} \leq (1+2\epsilon^{rep})q_{t+1}(S|m^t),$$

where the last inequality is true for $\epsilon^{rep} \leq 1/2$. ◀

Suppose that the labeled examples are sorted in some way and denote by $S^{\geq i}$ all the examples except the first $i-1$ examples.

▶ **Claim 33.** *If the hypotheses graph $(\mathcal{H},\mathcal{X},E)$ is* d*-mixing, then for any $\epsilon,\epsilon'>0$ except for a fraction of $\frac{1}{\epsilon^2}\cdot\frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}$ of the hypotheses $h\in\mathcal{H}$, for each $i\leq(1-\epsilon')|\mathcal{X}|$,*

$$\Pr(S^{\geq i}|h)\leq\left(1+\epsilon'+\frac{4\epsilon}{(\epsilon')^2}\right)\frac{|S^{\geq i}|}{|\mathcal{X}|}.$$

Proof. We will pick $\epsilon_1,\epsilon_2,\epsilon_3>0$ at the end. Divide all the labeled examples into $1/\epsilon_2$ consecutive equal parts, each of size $\epsilon_2|\mathcal{X}|$ (without loss of gnerality the integer $\epsilon_2|\mathcal{X}|$ divides $|\mathcal{X}|$). Focus for now on some part S. First we would like to show that for each part $S\subseteq\mathcal{X}$ most hypotheses h do not over-sample S, i.e.,

$$\Pr(S|h)\leq(1+\epsilon_1)\frac{|S|}{|\mathcal{X}|}.$$

Denote by $T\subseteq\mathcal{H}$ all the hypotheses $h\in\mathcal{H}$ such that $\Pr(S|h)>\frac{|S|}{|\mathcal{X}|}(1+\epsilon_1)$. Then $E(S,T)>\frac{|S|}{|\mathcal{X}|}(1+\epsilon_1)\frac{|\mathcal{X}|}{2}|T|$. From the d-mixing property we know that $E(S,T)\leq|S||T|/2+\mathrm{d}\sqrt{|S||T|}$. Combining these two inequalities we get that

$$\epsilon_1\frac{|S||T|}{2}<\mathrm{d}\sqrt{|S||T|}\Rightarrow|T|<\frac{4\mathrm{d}^2}{\epsilon_1^2|S|}=\frac{4\mathrm{d}^2}{\epsilon_1^2\epsilon_2|\mathcal{X}|}.$$

Denote by $Err\subseteq\mathcal{H}$ all the hypotheses that over-sample at least one part, i.e., hypothesis $h\notin Err$ if and only if for each of the $1/\epsilon_2$ parts, S, it holds that $\Pr(S|h)\leq(1+\epsilon_1)\frac{|S|}{|\mathcal{X}|}$. We can easily deduce, using a union bound, that the fraction of this set is at most $\frac{|Err|}{|\mathcal{H}|}\leq\frac{4\mathrm{d}^2}{\epsilon_1^2\epsilon_2^2|\mathcal{X}||\mathcal{H}|}$.

Let us go back to the expressions that we want to bound, namely $\Pr(S^{\geq i}|h)$ for each i. We will show that for each $h\notin Err$, and for each i, the probability

$$\Pr(S^{\geq i}|h)\leq(1+\epsilon_3)\frac{|S^{\geq i}|}{|\mathcal{X}|}. \tag{17}$$

For each i denote by i^* the largest index that is smaller than i and divides $\epsilon_2|\mathcal{X}|$. We have that $\Pr(x|h)\leq\frac{2}{|\mathcal{X}|}$ for each labeled example x and hypothesis h, thus $\Pr(S^{\geq i}\setminus S^{\geq i^*}|h)\leq2\epsilon_2$. Hence, the LHS of Inequality (17) is bounded by

$$\Pr(S^{\geq i}|h)\quad\leq\quad\Pr(S^{\geq i^*}|h)+2\epsilon_2\leq(1+\epsilon_1)\frac{|S^{\geq i}|}{|\mathcal{X}|}+2\epsilon_2,$$

So we need to make sure that

$$(1+\epsilon_1)\frac{|S^{\geq i}|}{|\mathcal{X}|}+2\epsilon_2\leq(1+\epsilon_3)\frac{|S^{\geq i}|}{|\mathcal{X}|},$$

which will happen only if $\epsilon_1\frac{|S^{\geq i}|}{|\mathcal{X}|}+2\epsilon_2\leq\epsilon_3\frac{|S^{\geq i}|}{|\mathcal{X}|}$, or equivalently $\frac{2\epsilon_2}{\epsilon_3-\epsilon_1}\leq\frac{|S^{\geq i}|}{|\mathcal{X}|}$ (assuming $\epsilon_3>\epsilon_1$ as we will choose later). Thus, except for a fraction of $\frac{4\mathrm{d}^2}{\epsilon_1^2\epsilon_2^2|\mathcal{X}||\mathcal{H}|}$ hypotheses $h\in\mathcal{H}$ for each $i\leq(1-\frac{2\epsilon_2}{\epsilon_3-\epsilon_1})|\mathcal{X}|$,

$$\Pr(S^{\geq i}|h)\leq(1+\epsilon_3)\frac{|S^{\geq i}|}{|\mathcal{X}|}.$$

Choose $\epsilon_1=\epsilon'$ and $\epsilon_2=\frac{2\epsilon}{\epsilon_1}$ and $\epsilon_3=\epsilon_1+\frac{2\epsilon_2}{\epsilon_1}$. ◀

▶ **Claim 34.** *For any $0<x\leq1/16$ it holds that $2^{\log_x(1-x)}\leq1+\sqrt{x}$.*

▶ **Fact 35.** *For any $x>-1$ it holds that $\frac{x}{1+x}\leq\ln(1+x)\leq x$.*

References

1. P. Beame, S. O. Gharan, and X. Yang. Time-space tradeoffs for learning from small test spaces: Learning low degree polynomial functions. Technical report, ECCC, 2017.
2. B. Chazelle. *The Discrepancy Method: Randomness and Complexity*. Randomness and Complexity. Cambridge University Press, 2000.
3. S. Garg, R. Raz, and A. Tal. Extractor-based time-space lower bounds for learning. Technical report, ECCC, 2017.
4. G. Kol, R. Raz, and A. Tal. Time-space hardness of learning sparse parities. In *Proc. 49th ACM Symp. on Theory of Computing*, 2017.
5. M. Krivelevich and B. Sudakov. Pseudo-random graphs. In *More sets, graphs and numbers*, pages 199–262. Springer, 2006.
6. D. Moshkovitz and M. Moshkovitz. Mixing implies lower bounds for space bounded learning. Technical report, ECCC Report TR17-017, 2017.
7. D. Moshkovitz and M. Moshkovitz. Mixing implies strong lower bounds for space bounded learning. Technical Report TR17-116, ECCC, 2017.
8. R. Raz. Fast learning requires good memory: A time-space lower bound for parity learning. In *Proc. 57th IEEE Symp. on Foundations of Computer Science*, 2016.
9. R. Raz. A time-space lower bound for a large class of learning problems. In *Proc. 58th IEEE Symp. on Foundations of Computer Science*, 2017.