Electronic Colloquium on Computational Complexity, Report No. 116 (2017)

Mixing Implies Strong Lower Bounds for Space Bounded Learning

Dana Moshkovitz* Michal Moshkovitz†

July 4, 2017

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 $2^{\Omega(\log^2|\mathcal{H}|)}$ memory states unless the learner uses at least a large number of $|\mathcal{H}|^{\Omega(1)}$ labeled examples. In contrast, there is a learner that uses $2^{\Theta(\log|\mathcal{X}|\log|\mathcal{H}|)}$ memory states and only $\Theta(\log|\mathcal{H}|)$ labeled examples, and there is a learner that uses only $|\mathcal{H}|$ memory states but a large number $\Theta(|\mathcal{H}|\log|\mathcal{H}|)$ of labeled examples. Our work builds on a combinatorial framework we suggested in a previous work for proving lower bounds on space bounded learning. The strong lower bound is obtained by considering a new notion of pseudorandomness for a sequence of graphs that represents the learner.

[*] danama@cs.utexas.edu. Department of Computer Science, UT Austin. This material is based upon work supported by the National Science Foundation under grants number 1218547 and 1648712.

[†] michal.moshkovitz@mail.huji.ac.il. The Edmond and Lily Safra Center for Brain Sciences, Hebrew University. 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.

ISSN 1433-8092

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 $|\mathcal{X}|^{\Theta(\log|\mathcal{H}|)}$ memory states. It is also possible to learn $\mathcal{H}$ using many fewer memory states: 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 $|\mathcal{H}|$ memory states but requires an extravagant number $\Theta(|\mathcal{H}|\log|\mathcal{H}|)$ of labeled examples. A natural question is whether one can learn with *both* $\ll |\mathcal{X}|^{\Theta(\log|\mathcal{H}|)}$ memory states and $\ll |\mathcal{H}|$ labeled examples.

Perhaps surprisingly, Raz [6] 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 $|\mathcal{X}|^{\Omega(\log|\mathcal{H}|)} = 2^{\Omega(n^2)}$ memory states 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 [5]. Many other applications follow using the large body of research on combinatorial mixing (see, e.g., [1]). 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., [5]). 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,* $m := \frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}$ *and* $|\mathcal{H}|$ *are at least some constants, then any learning algorithm that outputs the underlying hypothesis with probability at least* $m^{-\Theta(1)}$ *must use at least* $2^{\Omega(\log^2 m)}$ *memory states or* $m^{\Omega(1)}$ *labeled examples.*

[1] Kol, Raz and Tal [3] generalized Raz's work to parities on l variables out of n, showing that either $2^{\Omega(nl)}$ memory states or $2^{\Omega(l)}$ examples are needed, and for $l \leq n^{0.9}$, either $2^{\Omega(nl^{0.99})}$ memory states 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 [3]. (2) The work [3] 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.

A similar theorem holds if the learner only *approximately* learns the underlying hypothesis [5].

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 [5]. In [5] 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 $|\mathcal{H}|^{\Omega(1)}$). Independently of those two works (the current work and [5]) Raz [7] showed a result similar to the one in the current paper, relying on a spectral mixing condition instead of a combinatorial mixing condition.

1.2 Key Ideas

A key object in the combinatorial framework of [5] is the *knowledge graph* of the algorithm at various time steps. The knowledge graph is a bipartite graph, where one side corresponds to memory states and the other side corresponds to the possible hypotheses. 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. For every memory state, its neighborhood in the knowledge graph 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.

The paper [5] defines the notion of *certainty.* Low certainty implies that there are many plausible hypotheses for a typical memory state, whereas high certainty implies that one can guess the underlying hypothesis with good probability given the memory state. The combinatorial framework centers around bounding the certainty at every step of the algorithm, showing that it can only go up sufficiently after many time steps (i.e., after seeing many labeled examples).

Towards the goal of bounding the certainty, the work [5] 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. The intuition is that thanks to the bounded space very little information can be known about the underlying hypothesis provided that only bounded information is known about the memory state. The exact notion of "psuedorandom" is similar to an extractor property, except that the knowledge graph is determined by the algorithm and may be highly irregular, so we cannot use the standard definitions of extractor and min-entropy, but rather more general notions that we develop.

Unfortunately, an extractor-like property no longer holds when we wish to rule out learners that use, say, $|\mathcal{X}|^2$ memory states, let alone when the number of memory states is $|\mathcal{X}|^{\Theta(\log|\mathcal{H}|)}$. In this case the algorithm may store a whole labeled example in memory, and a large set of memory states may only span hypotheses consistent with that example (about half of the hypotheses). In order to handle this case, we introduce a new notion of pseudorandomness for the knowledge graph. The notion is a suitably *enhanced* sampler with multiplicative error. We describe it next without the modifications required due to irregularity.

We say that the knowledge graph is a *sampler with multiplicative error* L if the following property holds: For every probability distribution p with sufficient min-entropy k over the memory states $\mathcal{M}$, for every sufficiently large $H \subseteq \mathcal{H}$, the set H is sampled according to its fraction, up to a multiplicative error L, i.e.:

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

where $E(\cdot,\cdot)$ denotes the set of edges between given memories and hypotheses in the knowledge graph. For intuition, consider the case where the algorithm stores some of the labeled examples in memory. While the algorithm knows certain information about the hypothesis (it is consistent with the stored examples), the amount of information is limited. Hence, the probability of certain events H may grow, but not too much. Indeed, one can show that if the knowledge graph is a sampler with low multiplicative error throughout the execution of the algorithm, then the required lower bound follows. However, to prove that the sampler invariant is preserved, we need an enhanced property that we discuss next.

In our enhanced notion, for every probability distribution p over memories, we wish to benefit from memory states m whose probability is much lower than 2^{-k}, where k is p's min-entropy. The intuition is that such memories can be thought of as coming from a much higher "entropy level", and hence should give rise to a much lower multiplicative factor than the rest of the memories. Formally, for every $m \in \mathcal{M}$, denote $p(m) = 2^{-k} \cdot \gamma^{k_m}$ where γ is a parameter related to the mixing of the hypotheses graph and $k_m \geq 0$. We'd like the following condition to hold:

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

We show that the knowledge graph is a sampler with low multiplicative error 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. Moreover, roughly speaking, less likely transitions from time t to time $t+1$ may give a lot of information about the underlying hypothesis. The enhanced notion guarantees that even after taking the new information into account, we still have a sampler with multiplicative error (the actual analysis is quite involved, partly because it takes irregularity into account). The enhanced sampler notion might be of independent interest for pseudorandomness.

2 Preliminaries

$\log(\cdot)$ always means $\log_2(\cdot)$. The following claims were proven in [5]:

Claim 2. *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 3 (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 4 (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 5. *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 6 (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 7 (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 8 (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 [5]. □

3 The Low Certainty Framework

In this section we will summarize the main components of the combinatorial framework presented in our earlier work [5].

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 9 (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 [5] 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 [4]).

We can rephrase Claim 8 for the hypotheses graph and get

Proposition 10. *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 6).

Definition 11 (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 12. *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 13 (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 K-expander

To achieve the new results we need a different definition of pseudorandomness of the knowledge graph. This definition will be discussed in Section 4.

3.5 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.

In this section we define the key notion of *certainty*. The certainty of a memory captures the information it has on the underlying hypothesis, whereas the certainty of an hypothesis captures the information it has on the memory state to be reached assuming the hypothesis was picked. We further define the average certainty over all memories or hypotheses. We will consider memories or hypotheses that are "certain above average" as bad. An algorithm that successfully learns $\mathcal{H}$ will transform from having low average certainty at the initial stage to having high average certainty by its termination. Our argument will show that this increase in average certainty must take a long time.

First, we define the certainty of memories.

Definition 14 (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*

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

If, for example, all the hypotheses could have caused the algorithm to reach m with the same probability, then m's certainty is $\sum_h q_t(h|m)^2 = \frac{1}{|\mathcal{H}|}$ (e.g., this holds for the initial memory). If, on the other hand, given a memory m there is only one hypothesis h^* that caused the algorithm to reach this memory m then m's certainty is $\sum_h q_t(h|m)^2 = 1$.

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.

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^c_M (and additional memories, as will be defined in later sections). The next claim proves that removing bad memories does not reduce the weight too much. The following claims are proved in [5].

Claim 15. *For any $c > 0$ and time t, $q_t(Bad^c_M) \leq 1/c$*

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

Claim 16. *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 17. *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

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

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 $q_t(h) = 0$. Next claim proves that H_t is large.

Claim 18. *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 in Section 7 we choose $c \sim \log \frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}$.

In the next claim we will show that small certainty, small fraction of edges removed and $q_t(M_t) \approx 1$ imply that learning fails after t steps.

Claim 19. *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*

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

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

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

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

3.6 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 20. *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 [5] a weaker notion of *NRep* with some specific constant ϵ^{rep} was used.

Claim 21. *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 3.6.1.

3.6.1 Auxiliary Claims

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

Claim 22. *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 3 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 22, 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 4 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 23. *We say that a left regular bipartite graph (A, B, E) is $(\beta, \epsilon)-$weighted-expander with weights $w_1, \dots, w_{|A|}$, $\sum_i w_i = 1$, $\forall i, w_i \geq 0$, 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 24. *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 $w_1, \dots, w_{|\mathcal{H}|}$.*

You can find the proof of this claim in [5]. Next we will prove our main claim in this section.

Proof. (of Claim 21) 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 24 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 24. So again we conclude that $|T_2|<\beta|\mathcal{X}|$. □

3.7 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 [5].

4 Knowledge Graph Remains K-Expander

In this section we define a pseudorandom property, called *K-expander*, for the knowledge graph. We then prove that a K-expander remains a K-expander even in the face of a new labeled example, provided that the certainty is low and the hypotheses graph is mixing.

Definition 25 (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 $p(m) \le \frac{q(m)}{\beta}$ and if $p(m) > 0$ then $p(m) \ge \frac{q(m)}{\beta} \cdot \gamma$.*

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 26 (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 27 (K-expander)**.** *We say that the knowledge graph G'_t is an $(\alpha, \beta, \ell, \gamma_0, k) -$ K-expander if for every $H \subseteq \mathcal{H}$ with $|H| \ge \alpha|\mathcal{H}|$ and every (β, γ_0^k)-enlarging distribution p it holds that*

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

The usual definition of sampler with multiplicative error is

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

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 28. *Let $t \ge 1$. Assume that the following conditions hold:*

1. *The hypotheses graph is* d*-mixing.*
2. *The graph G'_t is an $(\alpha', \beta', \ell, \gamma_0, k) -$ K-expander.*
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, $cer(m^t) \le c \cdot cer^t(M_t)$ and $cer^t(M_t) \le c/|\mathcal{H}|$.*
5. *$\beta' \ge \gamma_0^{k-1}$ and $\alpha' \ge 2^{k+2}\sqrt{\gamma_0} + 2^{k+2} \cdot c \cdot \sqrt{\frac{16}{\gamma_0^{11}} \cdot \frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}}$.*
6. *$\epsilon^{rep} \le 1/2$, and $\gamma_0 \le 1/16$.*

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

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

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

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

$$\text{(Claim 5)} = \sum_{\substack{m^{t+1}\in M_{t+1}\\ m^t\in M_t}} q_{t+1}(H|m^t, S^{m^t,m^{t+1}})\frac{q_{t+1}(m^t, S^{m^t,m^{t+1}})}{q_{t+1}(m^{t+1})}p(m^{t+1})2^{\log_{\gamma_0} \frac{\beta'}{\beta'_{m^{t+1}}}}$$

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

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

$$\text{(Claim 4)} = \sum_{\substack{m^{t+1}\in M_{t+1}\\ m^t\in M_t, h\in H}} q_t(h|m^t)\frac{q_{t+1}(S^{m^t,m^{t+1}}|m^t,h)}{q_{t+1}(S^{m^t,m^{t+1}}|m^t)}\frac{q_t(m^t)q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}}2^{\log_{\gamma_0} \frac{\beta'}{\beta'_{m^{t+1}}}}$$

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

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,

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

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

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

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

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 K-expander 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

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

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*:

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

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

$$\begin{aligned}
&\sum_{m^t \in Low, h \in H} q_t(h|m^t) q_t(m^t) \sum_{m^{t+1} \in M_{t+1}} \frac{\Pr(S^{m^t,m^{t+1}}|h)}{\beta'_{m^{t+1}}} \cdot 2^k \\
\text{(see Claim 30)} \leq &\sum_{m^t \in Low, h \in H} q_t(h|m^t) q_t(m^t) \sum_{m^{t+1} \in M_{t+1}} \frac{2(1+2\epsilon^{rep}) q_{t+1}(S^{m^t,m^{t+1}}|m^t)}{\beta'_{m^{t+1}}} \cdot 2^k \\
\text{(definition of } p(m^t)) = &\sum_{m^t \in Low} q_t(H|m^t) 2^{k+1} (1+2\epsilon^{rep}) p(m^t) \\
(q_t(H|m^t) \leq 1, \epsilon^{rep} \leq 1/2) \leq &\, 2^{k+2} \sum_{m^t \in Low} p(m^t) \\
\text{(see Inequality (5))} \leq &\, 2^{k+2} \gamma_0 \qquad (7)
\end{aligned}$$

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

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

We define a distribution p' over memories at time t: if $m^t \in Low$ then $p'(m^t) = 0$, else $p'(m^t) = p(m^t)/(1-s)$. For convenience, we henceforth denote $(1-s)\beta'_{m^{t+1}}$ by $\beta'_{m^{t+1}}$.

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

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

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.

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

Let us start with bounding the first term in Equation (9). From Claim 31, 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 30 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

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

Let us now bound the second term in Expression (9). For each $i > i_{(m^t)}$ we use the simple bound given in Claim 30:

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

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

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

Which is equal to

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

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 $\frac{q_{t+1}(m^{t+1})}{\beta'_{m^{t+1}}} = p(m^{t+1}) \ge \frac{q_{t+1}(m^{t+1})\gamma_0^k}{\beta'}$, hence $2^{\log_{\gamma_0}\frac{\beta'}{\beta'_{m^{t+1}}}} \le 2^k$. 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 30 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 2 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 K-expander 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

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

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

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

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 29) 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 29), 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 33 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 33 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 K-expander 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 $\sqrt{\epsilon'} \le \gamma_0$) and $\epsilon = \gamma_0^5/4$. From the assumption in the claim we know that $\alpha'\sqrt{\gamma_0} \ge 2^{k+2}\gamma_0 + 2^{k+2} \cdot c \cdot \sqrt{\frac{16}{\gamma_0^{10}} \cdot \frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}}$. Hence, Expression (1) is at most

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

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

4.1 Auxiliary Claims

Claim 29. *For any $\epsilon \le 1/2$, the function $f(x) = 2^{\log_\epsilon \frac{1}{x}}$ for $x > 0$ is monotonically increasing and concave.*

Proof. Note that

$$f(x) = 2^{\frac{-\log x}{\log \epsilon}} = x^{\frac{1}{\log 1/\epsilon}}$$

The first derivative of f is equal to

$$f'(x) = \left(\frac{1}{\log 1/\epsilon}\right) \cdot x^{\frac{1}{\log 1/\epsilon}-1}.$$

The second derivative of f is equal to

$$f''(x) = \left(\frac{1}{\log 1/\epsilon}\right) \cdot \left(\frac{1}{\log 1/\epsilon} - 1\right) \cdot x^{\frac{1}{\log 1/\epsilon}-2}.$$

The terms $x^{\frac{1}{\log 1/\epsilon}-1}$ and $x^{\frac{1}{\log 1/\epsilon}-2}$ are both positive since $x > 0$. Since $\epsilon \le 1/2 < 1$ we know that $(\log 1/\epsilon)^{-1} > 0$. The term $\left(\frac{1}{\log 1/\epsilon} - 1\right)$ is smaller than 0 since $\frac{1}{\log 1/\epsilon} < 1 \Leftrightarrow 2 < 1/\epsilon \Leftarrow \epsilon \le 1/2$. □

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 30. *Let $S \subseteq \mathcal{X}$. Let $h \in \mathcal{H}$.*

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

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 20) to deduce that

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

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

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 31. *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 $\Pr(S^{\geq i} \backslash S^{\geq i^*}|h) \leq 2\epsilon_2$. 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 32. *For any $0 < x \leq 1/16$ it holds that*

$$2^{\log_x(1-x)} \leq 1 + \sqrt{x}.$$

Proof.

$$\begin{aligned}
& & & 2^{\log_x(1-x)} \leq 1+\sqrt{x} \\
& & \Leftrightarrow \quad & \log_x(1-x) \leq \log_2(1+\sqrt{x}) \\
& & \Leftrightarrow \quad & \frac{\ln(1-x)}{\ln x} \leq \frac{\ln(1+\sqrt{x})}{\ln 2} \\
& (x<1) & \Leftrightarrow \quad & \ln(1-x) \geq \frac{\ln x \cdot \ln(1+\sqrt{x})}{\ln 2} \\
& \left(\text{Fact 33} : \frac{-x}{1-x} \leq \ln(1-x)\right) & \Leftarrow \quad & \frac{-x}{1-x} \geq \frac{\ln x \cdot \ln(1+\sqrt{x})}{\ln 2} \\
& (\text{Fact 33} : x-1 \geq \ln(x)) & \Leftarrow \quad & \frac{-x}{1-x} \geq \frac{(x-1)\cdot \ln(1+\sqrt{x})}{\ln 2} \\
& & \Leftrightarrow \quad & \frac{x}{1-x} \leq \frac{(1-x)\cdot \ln(1+\sqrt{x})}{\ln 2} \\
& \left(\text{Fact 33} : \frac{\sqrt{x}}{1+\sqrt{x}} \leq \ln(1+\sqrt{x})\right) & \Leftarrow \quad & x\ln 2 \leq (1-x)^2 \cdot \frac{\sqrt{x}}{1+\sqrt{x}} \\
& & \Leftrightarrow \quad & \sqrt{x}(1+\sqrt{x})\ln 2 \leq (1-x)^2 \\
& (x \leq 1/16) & \Leftarrow \quad & \sqrt{x} \leq (1-x)^2,
\end{aligned}$$

and the last inequality is true for $x \leq 1/16$. □

Fact 33. *For any $x > -1$ it holds that*

$$\frac{x}{1+x} \leq \ln(1+x) \leq x$$

Proof. Our starting point is the known inequality

$$1 + x \leq e^x \tag{18}$$

which is true for any x. This immediately proves the second inequality in the fact. To prove the first inequality, substitute $x = -\ln(1+y)$ in Inequality (18) and get

$$1 - \ln(1+y) \leq \frac{1}{1+y}.$$

Equivalently,

$$\frac{y}{1+y} \leq \ln(1+y)$$

□

5 Heavy Sourced Memories

We start by examining one possible step of the algorithm: when there is an abundance of examples $S \subseteq \mathcal{X}$ that lead from a memory m^t at time t to a memory m^{t+1} at time $t+1$. The algorithm can apply such a step, for example, to examine consistency with a specific hypothesis h. All the labeled examples that are consistent with h (there are $|\mathcal{X}|/2$ such labeled examples) will lead the algorithm to change the memory state from m^t to m^{t+1}.

Definition 34. *The set of heavy-sourced memories at time $t+1$ is defined as*

$$M_{t+1}^{heavy>b} = \{m^{t+1} | \exists m^t \in M_t \textit{ with at least } b|X| \textit{ labeled examples that lead to } m^{t+1}\}.$$

We will assume, without loss of generality, that m^{t+1} cannot be reached through other memories (otherwise, make a few copies of m^{t+1}; we will make this argument formal in Section 7). Under this assumption it makes sense to identify – as we will do later – a memory m^{t+1} with a pair (m^t, S) that lead to it.

We would like to show that the certainty does not increase much as a result of heavy steps. The intuition is that if there is low certainty at m^t, then the mixing of the hypotheses graph ensures that S reveals very little information on which of the possible hypotheses is the underlying one. The bound on the certainty at time $t+1$ as a function of the certainty at time t is shown in Claim 35 and in Claim 36. Claim 35 gives an expression for $cer^{t+1}(M_{t+1}^{heavy>b})$. To understand this expression, notice that a small variant of Claim 17 is the following equality

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

Claim 35. *If $|H_{t+1}| \geq |H_t|(1-1/c)$, and $c \geq 2$ then for any set M of memories at time $t+1$ and any weighted vector w (i.e., $\forall i, w_i \in [0,1]$) it holds that $cer_w^{t+1}(M_{t+1}^{heavy>b} \cap M)$ is at most*

$$\left(1 + \frac{2}{c}\right) \sum_{\substack{(m^t,S) \in M_{t+1}^{heavy>b} \cap M \\ h \in \mathcal{H}}} q_t(h) q_t(h|m^t) q_t(m^t|h) w_{(m^t,S)} \Pr(S|h) \frac{\Pr(S|h)}{\sum_{h'} q_t(h'|m^t) \Pr(S|h')}$$

Proof. Let us start with rewriting $q_{t+1}(h|m^{t+1})$, for some $m^{t+1} \in M_{t+1}^{heavy>b}$ that corresponds to the pair (m^t, S)

$$\begin{aligned}
(\star) \qquad q_{t+1}(h|m^{t+1}) &= q_{t+1}(h|S, m^t) \\
\text{(using Claim 4)} &= q_{t+1}(S|h, m^t)\frac{q_t(h|m^t)}{q_{t+1}(S|m^t)} \\
\text{(using } \Pr(S|h, m^t) = \Pr(S|h), &= \frac{\Pr(S|h)q_t(h|m^t)}{\sum_{h'} q_{t+1}(S|m^t, h')q_t(h'|m^t)} \\
\text{and Claim 3)} & \\
&= \frac{\Pr(S|h)q_t(h|m^t)}{\sum_{h'} \Pr(S|h')q_t(h'|m^t)}
\end{aligned}$$

Note that

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

Use Claim 17 and equations $(\star), (\star, \star)$ to rewrite $cer_w^{t+1}(M_{t+1}^{heavy>b} \cap M)$

$$\begin{aligned}
& \sum_{h\in\mathcal{H}} q_{t+1}(h) \sum_{m^{t+1}\in M_{t+1}^{heavy>b}\cap M} w_{m^{t+1}} q_{t+1}(h|m^{t+1})q_{t+1}(m^{t+1}|h) \\
&= \sum_{h\in\mathcal{H}} q_{t+1}(h) \sum_{(m^t,S)\in M_{t+1}^{heavy>b}\cap M} w_{(m^t,S)} \frac{\Pr(S|h)q_t(h|m^t)}{\sum_{h'} \Pr(S|h')q_t(h'|m^t)} q_t(m^t|h)\Pr(S|h) \\
\text{(see below)} \quad &\leq \left(1+\frac{2}{c}\right) \sum_{h\in\mathcal{H}} q_t(h) \sum_{(m^t,S)\in M_{t+1}^{heavy>b}\cap M} w_{(m^t,S)} \frac{\Pr(S|h)q_t(h|m^t)}{\sum_{h'} \Pr(S|h')q_t(h'|m^t)} q_t(m^t|h)\Pr(S|h) \\
&= \left(1+\frac{2}{c}\right) \sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in\mathcal{H}}} q_t(h)q_t(h|m^t)q_t(m^t|h)w_{(m^t,S)}\Pr(S|h)\frac{\Pr(S|h)}{\sum_{h'} \Pr(S|h')q_t(h'|m^t)},
\end{aligned}$$

to understand why the inequality is true, notice that we have a sum of the form $\sum_{h\in\mathcal{H}} q_{t+1}(h)a_h$ for some value $a_h \geq 0$, which is equal (by the definition of $q_t(h)$) to

$$\begin{aligned}
\frac{1}{|H_{t+1}|}\sum_{h\in H_{t+1}} a_h &\leq \frac{1}{|H_t|(1-1/c)}\sum_{h\in H_{t+1}} a_h \\
\text{(for } c \geq 2) &\leq \left(1+\frac{2}{c}\right)\frac{1}{|H_t|}\sum_{h\in H_{t+1}} a_h \\
(H_{t+1} \subseteq H_t) &\leq \left(1+\frac{2}{c}\right)\frac{1}{|H_t|}\sum_{h\in H_t} a_h \\
&= \left(1+\frac{2}{c}\right)\sum_{h\in\mathcal{H}} q_t(h)a_h
\end{aligned}$$

□

The next claim shows that certainty does not increase much in the case of heavy sourced memories.

Claim 36. *If the hypotheses graph is an (ϵ, ϵ')-sampler, $c \geq 4$, $|H_t| \geq |\mathcal{H}|/3$, $|H_{t+1}| \geq |H_t|(1 - 1/c)$, $cer^t(M_t) \leq \frac{c}{|\mathcal{H}|}$, and for each $m \in M_t, h \in H_t$, it holds that $q_t(h|m) \leq a \cdot cer^t(M_t)$, and*

$$b \geq \max(5\epsilon c + 2c^2\sqrt{\epsilon'}, 4a^2\epsilon' c^2 + \epsilon),$$

then for any set of memories M at time $t+1$ and any weight w it holds that

$$cer_w^{t+1}(M_{t+1}^{heavy>b} \cap M) \leq \left[\left(1 + \frac{4}{c}\right) \sum_{(m^t,S) \in M_{t+1}^{heavy>b} \cap M} cer^t(m^t)\frac{|S|}{|\mathcal{X}|} w_{(m^t,S)}\right] + \left[\frac{2}{c} \cdot cer^t(M_t)\right]$$

Proof. For each subset of labeled examples $S \subseteq \mathcal{X}$ define $Err(S) \subseteq \mathcal{H}$ as the set of all hypotheses that do not sample S correctly, i.e., if $h \in Err(S)$, then $\left|\Pr(S|h) - \frac{|S|}{|\mathcal{X}|}\right| > \epsilon$. From the sampler property of the hypotheses graph (see Definition 7) we know that for every $S \subseteq \mathcal{X}$, $|Err(S)| \leq \epsilon' |\mathcal{H}|$.

According to Claim 35, $cer_w^{t+1}(M_{t+1}^{heavy>b} \cap M)$ is at most $(\star)$

$$\left(1 + \frac{2}{c}\right) \sum_{\substack{(m^t,S) \in M_{t+1}^{heavy>b} \cap M \\ h \in \mathcal{H}}} q_t(h) q_t(h|m^t) q_t(m^t|h) w_{(m^t,S)} \Pr(S|h) \frac{\Pr(S|h)}{\sum_{h'} q_t(h'|m^t)\Pr(S|h')}$$

The denominator can be lower bounded using the sampler property of the hypotheses graph as follows

$$\begin{aligned}
\sum_{h'} q_t(h'|m^t)\Pr(S|h') &\geq \sum_{h' \notin Err(S)} q_t(h'|m^t)\Pr(S|h') \\
&\geq \left(\frac{|S|}{|\mathcal{X}|} - \epsilon\right) \sum_{h' \notin Err(S)} q_t(h'|m^t) \\
\text{(see below)} &\geq \left(\frac{|S|}{|\mathcal{X}|} - \epsilon\right)(1 - \epsilon''),
\end{aligned}$$

where in the last inequality we used Claim 2 with $\epsilon'' := \sqrt{\epsilon' |\mathcal{H}| c \cdot cer^t(M_t)}$ and the distribution $q_t(\cdot|m^t)$ we also used the fact that since $m^t \notin Bad_{M_t}$ we know that $\sum_h q(h|m^t)^2 \leq c \cdot cer^t(M_t)$. From the assumption in the claim we know that $cer^t(M_t) \leq \frac{c}{|\mathcal{H}|}$, this implies that $\epsilon'' \leq c\sqrt{\epsilon'}$.

Consider two cases:

Case 1: If $h \notin Err(S)$, then $\Pr(S|h) \le \frac{|S|}{|\mathcal{X}|} + \epsilon$. Thus,

$$\begin{aligned}
\frac{\Pr(S|h)}{\sum_{h'} q_t(h'|m^t)\Pr(S|h')} &\le \frac{\frac{|S|}{|\mathcal{X}|} + \epsilon}{\left(\frac{|S|}{|\mathcal{X}|} - \epsilon\right)(1-\epsilon'')} \\
&\le 1 + \frac{2\epsilon + \epsilon''}{\left(\frac{|S|}{|\mathcal{X}|} - \epsilon\right)(1-\epsilon'')} \\
\text{(using } |S|/|\mathcal{X}| \ge b) &\le 1 + \frac{2\epsilon + \epsilon''}{(b-\epsilon)(1-\epsilon'')}
\end{aligned}$$

Case 2: If $h \in Err(S)$, then we use $\Pr(S|h) \le 1$ to bound

$$\frac{\Pr(S|h)}{\sum_{h'} q_t(h'|m^t)\Pr(S|h')} \le \frac{1}{(b-\epsilon)(1-\epsilon'')}.$$

We will show that

$$\sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in Err(S)}} q_t(h)q_t(h|m^t)q_t(m^t|h)w_{(m^t,S)}\Pr(S|h) \le 2a^2c\epsilon' \cdot cer^t(M_t)$$

The left hand side is at most

$$\begin{aligned}
\text{(see below)} &\le \sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in Err(S)\cap H_t}} q_t(h)q_t(h|m^t)q_t(m^t|h)2\frac{|S|}{|\mathcal{X}|} \\
\text{(Claim 16)} &\le \sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in Err(S)\cap H_t}} q_t(m^t)q_t(h|m^t)^2 \cdot 2\frac{|S|}{|\mathcal{X}|} \\
\text{(assumption in the claim)} &\le \sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in Err(S)\cap H_t}} q_t(m^t)(a\cdot cer^t(M_t))^2 \cdot 2\frac{|S|}{|\mathcal{X}|} \\
(cer^t(M_t) \le \frac{c}{|\mathcal{H}|}) &\le \sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in Err(S)}} q_t(m^t)\frac{a^2c}{|\mathcal{H}|}\cdot cer^t(M_t)\cdot 2\frac{|S|}{|\mathcal{X}|} \\
(|Err(S)| \le \epsilon'|\mathcal{H}|) &\le \sum_{(m^t,S)\in M_{t+1}^{heavy>b}\cap M} q_t(m^t)\frac{a^2c}{|\mathcal{H}|}\cdot cer^t(M_t)\cdot 2\frac{|S|}{|\mathcal{X}|}\cdot\epsilon'|\mathcal{H}| \\
&\le 2a^2c\epsilon'\cdot cer^t(M_t)\cdot\sum_{m^t\in M_t} q_t(m^t) \\
&\le 2a^2c\epsilon'\cdot cer^t(M_t)
\end{aligned}$$

The first inequality is true from the following reasons: 1. $w_i \leq 1$, for each i 2. for each $x \in \mathcal{X}$, $\Pr(x|h)$ is either 0 or $2/|\mathcal{X}|$ 3. if $h \notin H_t$ then $q_t(h) = 0$.

To sum up the two cases, Equation ($\star$) is at most

$$\left(1+\frac{2}{c}\right)\Bigg[\Bigg[\sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M\\ h\in\mathcal{H}}} q_t(h)q_t(h|m^t)q_t(m^t|h)w_{(m^t,S)}\left(\frac{|S|}{|\mathcal{X}|}+\epsilon\right)\left(1+\frac{2\epsilon+\epsilon''}{(b-\epsilon)(1-\epsilon'')}\right)\Bigg] + 2a^2c\epsilon'\cdot cer^t(M_t)\frac{1}{(b-\epsilon)\,(1-\epsilon'')}\Bigg]$$

Using Claim 17, (i.e., $cer^t(m^t) = \sum_{h\in\mathcal{H}} q_t(h)q_t(h|m^t)q_t(m^t|h)$), Equation ($\star$) is at most

$$\left(1+\frac{2}{c}\right)\Bigg[\Bigg[\sum_{(m^t,S)\in M_{t+1}^{heavy>b}\cap M} cer^t(m^t)w_{(m^t,S)}\frac{|S|}{|\mathcal{X}|}\left(1+\frac{\epsilon}{|S|/|\mathcal{X}|}\right)\left(1+\frac{2\epsilon+\epsilon''}{(b-\epsilon)(1-\epsilon'')}\right)\Bigg] + 2a^2c\epsilon'\cdot cer^t(M_t)\frac{1}{(b-\epsilon)\,(1-\epsilon'')}\Bigg]$$

The rest of the proof uses simple algebraic manipulations.

$$\begin{aligned}\left(1+\frac{2}{c}\right)\left(1+\frac{\epsilon}{|S|/|\mathcal{X}|}\right)\left(1+\frac{2\epsilon+\epsilon''}{(b-\epsilon)(1-\epsilon'')}\right) &\leq \left(1+\frac{2}{c}\right)\left(1+\frac{\epsilon}{b}\right)\left(1+\frac{2\epsilon+\epsilon''}{(b-\epsilon)(1-\epsilon'')}\right)\\ \text{(see Items (1), (2) below)} &\leq \left(1+\frac{2}{c}\right)\left(1+\frac{1}{5c}\right)\left(1+\frac{1}{c}\right)\\ \text{(see Item (3) below)} &\leq 1+\frac{4}{c}\end{aligned}$$

1. $5\epsilon c \leq b \Rightarrow \frac{\epsilon}{b} \leq \frac{1}{5c}$

2. We would like to bound $\frac{2\epsilon+\epsilon''}{(b-\epsilon)(1-\epsilon'')}$ by $\frac{1}{c}$. Recall $\epsilon'' \leq c\sqrt{\epsilon'}$. We have $5\epsilon c + 2c^2\sqrt{\epsilon'} \leq b \leq 1 \Rightarrow \epsilon'' \leq c\sqrt{\epsilon'} \leq 0.5 \Rightarrow \frac{1}{1-\epsilon''} \leq 2$. Thus, we would like to show the bound $4\epsilon c + 2\epsilon'' c \leq b - \epsilon$, so it is enough that $5\epsilon c + 2c^2\sqrt{\epsilon'} \leq b$, which is true by the assumption in the claim.

3. The expression $\left(1+\frac{2}{c}\right)\left(1+\frac{1}{5c}\right)\left(1+\frac{1}{c}\right)$ is equal to

$$\begin{aligned} & \left(1+\frac{1}{5c}+\frac{2}{c}+\frac{2}{5c^2}\right)\left(1+\frac{1}{c}\right) \\ = \; & 1+\frac{1}{5c}+\frac{2}{c}+\frac{2}{5c^2}+\frac{1}{c}+\frac{1}{5c^2}+\frac{2}{c^2}+\frac{2}{5c^3} \\ = \; & 1+\frac{16}{5c}+\frac{13}{5c^2}+\frac{2}{5c^3} \\ = \; & 1+\frac{16}{5c}+\frac{4}{c}\cdot\frac{1}{20c}\left(13+\frac{2}{c}\right) \\ (c\geq 4) \quad \leq \; & 1+\frac{4}{c} \end{aligned}$$

Let us move on to the second expression we would like to bound

$$\begin{aligned} & \left(1+\frac{2}{c}\right)2a^2c\epsilon'\frac{1}{(b-\epsilon)\left(1-\epsilon''\right)} \\ \text{(see Item 1 below)} \quad \leq \; & \left(1+\frac{2}{c}\right)\frac{1}{c} \\ \text{(see Item 2 below)} \quad \leq \; & \frac{2}{c} \end{aligned}$$

1. It suffices to show that $\frac{4a^2c\epsilon'}{b-\epsilon} \leq 1/c \Leftrightarrow 4a^2\epsilon'c^2+\epsilon \leq b$
2. $(1+2/c)1/c = 1/c+2/c^2$ and also $2/c^2 \leq 1/c$ for $2 \leq c$.

□

6 Many Sourced Memories

We would like to show that the certainty remains low in the case that a new memory m^{t+1} is reached by sufficiently large q_t-weight memories $\psi(m^{t+1}) = \{m_1^t, m_2^t, \ldots\}$ at time t and each such memory m_i^t is reached using exactly one representative labeled example x_i. Recall that representative examples were defined in Section 3.6.

We will assume, without loss of generality, that m^{t+1} cannot be reached from m^t using more than one example (otherwise, make a few copies of m^{t+1}; we will make this argument formal in Section 7). Under this assumption it makes sense to identify – as we will do later – a memory m^{t+1} with set of memory-(labeled-)example pairs $\{(m_i^t, x_i)\}$ that lead to it.

Definition 37. *The set of many-sourced memories at time $t+1$ is defined as*

$$\begin{aligned} M_{t+1}^{many>\beta,\epsilon^{rep}} \; = \; & \{m^{t+1} | \exists \textit{ memories } m_i^t \in M_t \textit{ with } \sum_i q_t(m_i^t) \geq \beta \\ & \textit{and labeled examples } x_i \notin NRep(m_i^t, \epsilon^{rep}) \textit{ that lead to } m^{t+1}\}. \end{aligned}$$

We will prove that the certainty remains low for many-sourced memories for β that will be chosen later. Here is an outline of the proof (the exact values of the constants are not important):

1. Recall from the K-expander property (that its preservation we proved in Claim 28) that for any large enough $H \subseteq \mathcal{H}$ it holds that

$$q_t(H|\psi(m^{t+1})) \leq \ell \frac{|H|}{|\mathcal{H}|}$$

   (also recall that $\psi(m^{t+1})$ are the memories at time t that lead to m^{t+1}.)

2. We will prove that for any $h \in \mathcal{H}$,

$$q_{t+1}(h|m^{t+1}) \leq 2(1+\epsilon^{rep})q_t(h|\psi(m^{t+1}))$$

   The intuition is that one labeled example gives about one bit of information on h and this changes the probability by about a factor of 2.

3. Putting together the first two steps we have that except for a small size set $T \subset \mathcal{H}$, for any other $h \in \mathcal{H}$,

$$q_{t+1}(h|m^{t+1}) \leq \frac{2(1+\epsilon^{rep})\ell}{|\mathcal{H}|}.$$

   Importantly, the bound does not not depend on t.

4. Then we will show that certainty remains low.

In step 2 we want to upper bound $q_{t+1}(h|m^{t+1})$. Let us start with investigating this term and writing it as a function of memories from time t.

Claim 38. *For any hypothesis h and a memory m^{t+1} that can be reached by the pairs $\{(m_i^t, S_i)\}$ it holds that*

$$q_{t+1}(h|m^{t+1}) = \frac{\sum_i \Pr(S_i|h)q_t(h|m_i^t)q_t(m_i^t)}{\sum_i q_{t+1}(S_i|m_i^t)q_t(m_i^t)}$$

Proof.

$$\begin{aligned}
q_{t+1}(h|m^{t+1}) &= q_{t+1}(h| \vee_i (m_i^t, S_i)) \\
\text{(Conditional probability dfn.)} &= \frac{q_{t+1}\big(h \wedge (\vee_i(m_i^t, S_i))\big)}{q_{t+1}(\vee_i(m_i^t, S_i))} \\
\text{(De Morgan's law)} &= \frac{q_{t+1}\big(\vee_i (h \wedge (m_i^t, S_i))\big)}{q_{t+1}(\vee_i(m_i^t, S_i))} \\
\text{(Disjoint events)} &= \frac{\sum_i q_{t+1}\big(h \wedge (m_i^t, S_i)\big)}{\sum_i q_{t+1}(m_i^t, S_i)} \\
\text{(Conditional probability dfn.)} &= \frac{\sum_i q_{t+1}(h|m_i^t, S_i) q_{t+1}(m_i^t, S_i)}{\sum_i q_{t+1}(S_i|m_i^t) q_t(m_i^t)} \\
\text{(Claim 4 \& } q_{t+1}(S_i|h, m_i^t) = \Pr(S_i|h)\text{)} &= \frac{\sum_i \Pr(S_i|h) \frac{q_t(h|m_i^t)}{q_{t+1}(S_i|m_i^t)} q_{t+1}(S_i|m_i^t) q_t(m_i^t)}{\sum_i q_{t+1}(S_i|m_i^t) q_t(m_i^t)} \\
&= \frac{\sum_i \Pr(S_i|h) q_t(h|m_i^t) q_t(m_i^t)}{\sum_i q_{t+1}(S_i|m_i^t) q_t(m_i^t)}
\end{aligned}$$

□

Now we are ready to prove step 2.

Claim 39. *If* $m^{t+1} \in M_{t+1}^{many>\beta,\epsilon^{rep}}$ *and* $\epsilon^{rep} \leq 1/2$ *then for any* $h \in \mathcal{H}$ *it holds that*

$$q_{t+1}(h|m^{t+1}) \leq 2(1 + 2\epsilon^{rep}) \cdot q_t(h|\psi(m^{t+1})).$$

Proof. We will use the fact that if $m^{t+1} \in M_{t+1}^{many>\beta,\epsilon^{rep}}$, then it can be reached exactly by the memory-(labeled-)example pairs $\{(m_i^t, x_i)\}$ where all memories m_i^t are different and for all i, $x_i \notin NRep(m_i)$.

From Claim 38 with $S_i = \{x_i\}$ for all i we know that

$$\begin{aligned}
q_{t+1}(h|m^{t+1}) &= \frac{\sum_i \Pr(x_i|h) q_t(h|m_i^t) q_t(m_i^t)}{\sum_i q_{t+1}(x_i|m_i^t) q_t(m_i^t)} \\
\text{(see below)} &\leq \frac{\sum_i \frac{2}{|\mathcal{X}|} q_t(h|m_i^t) q_t(m_i^t)}{\sum_i q_{t+1}(x_i|m_i^t) q_t(m_i^t)} \\
\text{(Definition 20)} &\leq \frac{\sum_i \frac{2}{|\mathcal{X}|} q_t(h|m_i^t) q_t(m_i^t)}{\sum_i \frac{1-\epsilon^{rep}}{|\mathcal{X}|} q_t(m_i^t)} \\
(\epsilon^{rep} \leq 1/2) &= 2(1 + 2\epsilon^{rep}) \cdot \frac{\sum_i q_t(h|m_i^t) q_t(m_i^t)}{\sum_i q_t(m_i^t)} \\
&= 2(1 + 2\epsilon^{rep}) \cdot \sum_i q_t(h|m_i^t) \frac{q_t(m_i^t)}{q_t(\psi(m_i^{t+1}))} \\
\text{(by Claim 5)} &= 2(1 + 2\epsilon^{rep}) \cdot q_t(h|\psi(m^{t+1}))
\end{aligned}$$

the first inequality is true since if x_i and h are consistent then $\Pr(x_i|h) = \frac{2}{|\mathcal{X}|}$, else $\Pr(x_i|h) = 0$. □

Let us move to step 3.

Claim 40. *If the graph G'_t is an $(\alpha', \beta', \ell, \gamma_0, k) - K$-expander, then for every memory $m^{t+1} \in M_{t+1}^{many>\beta', \epsilon^{rep}}$ there is a set $T \subset \mathcal{H}, |T| \leq \alpha'|\mathcal{H}|$, such that for any $h \notin T$ it holds that*

$$q_{t+1}(h|m^{t+1}) \leq \frac{2(1+\epsilon^{rep})\ell}{|\mathcal{H}|}.$$

Proof. Define $T = \{h | \frac{2(1+2\epsilon^{rep})\ell}{|\mathcal{H}|} < q_{t+1}(h|m^{t+1})\}$, then

$$2(1+2\epsilon^{rep})\ell\frac{|T|}{|\mathcal{H}|} < q_{t+1}(T|m^{t+1}),$$

From Claim 39 we know that for every $h \in \mathcal{H}$ it holds that

$$q_{t+1}(h|m^{t+1}) \leq 2(1+2\epsilon^{rep}) \cdot q_t(h|\psi(m^{t+1})).$$

The last two inequalities imply that

$$2(1+2\epsilon^{rep})\ell\frac{|T|}{|\mathcal{H}|} < 2(1+2\epsilon_{rep}) \cdot q_t(T|\psi(m^{t+1})).$$

Assume by contradiction that $|T| \geq \alpha'|\mathcal{H}|$, then from the K-expander property we know that $q_t(T|\psi(m^{t+1})) \leq \ell\frac{|T|}{|\mathcal{H}|}$. Putting the last two inequalities together leads to a contradiction. □

Let us move on and prove the 4 step in the outline. To this end, we first prove that *vertex contraction* can only reduce certainty, where contracting a few memories $m_1, \ldots, m_l$ in the knowledge graph into one means that all these l vertices are replaced by one vertex m and all the edges of the form (m_i, h) are now of the form (m, h). Notice that the number of edges remains the same. The reason we care about vertex contraction is that from the point of view of the memory m^{t+1} the vertices $\psi(m^{t+1})$ were contracted.

Claim 41. *If memories $m_1, \ldots, m_l$ have been contracted to a vertex m, then*

$$q_t(m)q_t(h|m)^2 \leq \sum_i q_t(m_i)q_t(h|m_i)^2$$

Proof.

$$\begin{aligned}
q_t(m)q_t(h|m)^2 &= q_t(m)q_t(h|m_1 \vee \ldots \vee m_l)^2 \\
\text{(using Claim 5)} &= q_t(m)\left(\sum_i q_t(h|m_i)\frac{q_t(m_i)}{q_t(m)}\right)^2 \\
\text{(by Jensen's inequality)} &\leq q_t(m)\sum_i \left(q_t(h|m_i)^2\frac{q_t(m_i)}{q_t(m)}\right) \\
&= \sum_i q_t(m_i)q_t(h|m_i)^2
\end{aligned}$$

□

Using Claim 16, the last claim imply the following

Corollary 42. *If memories $m_1, \ldots, m_l$ have been contracted to a vertex m, then*

$$q_t(h)q_t(h|m)q_t(m|h) \leq \sum_i q_t(h)q_t(h|m_i)q_t(m_i|h)$$

Claim 43. *If the hypotheses graph is an $(\alpha, \beta, \epsilon)-$H-expander, the graph G'_t is an $(\alpha', \beta', \ell, \gamma_0, k)-$ K-expander, $c \geq 45$, $cer^t(M_t) \leq \frac{c}{|\mathcal{H}|}$, $|H_{t+1}| \geq (1 - 1/c)|H_t|$, $3|H_t| \geq |\mathcal{H}|$, and for each $m \in M_t, h \in H_t$, it holds that $q_t(h|m) \leq a \cdot cer^t(M_t)$, then for any set of memories M at time $t+1$ and any weighted vector w (i.e., $\forall i, w_i \in [0,1]$) it holds that*

$$cer_w^{t+1}(M_{t+1}^{many>\beta',\epsilon^{rep}} \cap M) \leq \left[\frac{2(1+2\epsilon^{rep})\ell}{|\mathcal{H}|} \cdot \sum_{m \in M_{t+1}^{many>\beta',\epsilon^{rep}} \cap M} q_{t+1}(m)w_m\right] + 10\alpha' ca^2 cer^t(M_t)$$

Proof. Using Claim 40 for every memory $m^{t+1} \in M_{t+1}^{many>\beta',\epsilon^{rep}}$ there is a set $T_{m^{t+1}} \subset \mathcal{H}, |T_{m^{t+1}}| \leq \alpha'|\mathcal{H}|$, such that for any $h \notin T_{m^{t+1}}$ it holds that

$$q_{t+1}(h|m^{t+1}) \leq \frac{2(1+2\epsilon^{rep})\ell}{|\mathcal{H}|}.$$

Using Claim 17,

$$\begin{aligned}
cer_w^{t+1}(M_{t+1}^{many>\beta',\epsilon^{rep}} \cap M) &= \sum_{\substack{m^{t+1} \in M_{t+1}^{many>\beta',\epsilon^{rep}} \cap M \\ h \in \mathcal{H}}} q_{t+1}(h)q_{t+1}(h|m^{t+1})q_{t+1}(m^{t+1}|h)w_{m^{t+1}} \\
&= \sum_{\substack{m^{t+1} \in M_{t+1}^{many>\beta',\epsilon^{rep}} \cap M \\ h \notin T_{m^{t+1}}}} q_{t+1}(h)q_{t+1}(h|m^{t+1})q_{t+1}(m^{t+1}|h)w_{m^{t+1}} + \\
&\quad \sum_{\substack{m^{t+1} \in M_{t+1}^{many>\beta',\epsilon^{rep}} \cap M \\ h \in T_{m^{t+1}}}} q_{t+1}(h)q_{t+1}(h|m^{t+1})q_{t+1}(m^{t+1}|h)w_{m^{t+1}}
\end{aligned}$$

The sum over $h \notin T_{m^{t+1}}$ is at most

$$\begin{aligned}
& \sum_{\substack{m^{t+1} \in M_{t+1}^{many>\beta',\epsilon^{rep}} \cap M \\ h \notin T_{m^{t+1}}}} q_{t+1}(h) \cdot \frac{2(1+2\epsilon^{rep})\ell}{|\mathcal{H}|} \cdot q_{t+1}(m^{t+1}|h) w_{m^{t+1}} \\
\leq \quad & \frac{2(1+2\epsilon^{rep})\ell}{|\mathcal{H}|} \cdot \sum_{m^{t+1} \in M_{t+1}^{many>\beta',\epsilon^{rep}} \cap M} w_{m^{t+1}} \sum_{h \in \mathcal{H}} q_{t+1}(h) q_{t+1}(m^{t+1}|h) \\
= \quad & \frac{2(1+2\epsilon^{rep})\ell}{|\mathcal{H}|} \cdot \sum_{m^{t+1} \in M_{t+1}^{many>\beta',\epsilon^{rep}} \cap M} q_{t+1}(m^{t+1}) w_{m^{t+1}}
\end{aligned}$$

Let us focus on the sum over $h \in T_{m^{t+1}}$. From Claim 39 we know that

$$q_{t+1}(h|m^{t+1}) \leq 2(1+2\epsilon^{rep}) q_t(h|\psi(m^{t+1})) \quad (\star)$$

We can also upper bound the term

$$\begin{aligned}
q_{t+1}(m^{t+1}|h) &= q_{t+1}(\vee_i (m_i^t, x_i)|h) \\
&= \sum_i q_{t+1}(m_i^t, x_i|h) \\
&= \sum_i q_t(m_i^t|h) \Pr(x_i|h) \\
\text{(see below)} \quad &\leq \sum_i q_t(m_i^t|h) \frac{2}{|\mathcal{X}|} \\
&= \frac{2}{|\mathcal{X}|} q_t(\psi(m^{t+1})|h) \quad (\star\star)
\end{aligned}$$

where the inequality is true since if $I_{(x,h)\in E}$ then $\Pr(x|h) = 2/|\mathcal{X}|$, else $\Pr(x|h) = 0$. Thus, from Equation $(\star), (\star\star)$ (and using $\forall_i w_i \in [0,1]$)

$$\begin{aligned}
&\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta',\epsilon^{rep}}\cap M\\ h\in T_{m^{t+1}}}} && q_{t+1}(h)q_{t+1}(h|m^{t+1})q_{t+1}(m^{t+1}|h)\\
& &\leq\;& \frac{4(1+2\epsilon^{rep})}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta',\epsilon^{rep}}\cap M\\ h\in T_{m^{t+1}}}} q_{t+1}(h)q_t(h|\psi(m^{t+1}))q_t(\psi(m^{t+1})|h)\\
&\text{(see below)} &\leq\;& \frac{10}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta',\epsilon^{rep}}\cap M\\ h\in T_{m^{t+1}}\cap H_t}} q_t(h)q_t(h|\psi(m^{t+1}))q_t(\psi(m^{t+1})|h)\\
&\text{(using Claim 42)} &\leq\;& \frac{10}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta',\epsilon^{rep}}\cap M\\ h\in T_{m^{t+1}}\cap H_t\\ m^t\in\psi(m^{t+1})}} q_t(h)q_t(h|m^t)q_t(m^t|h)\\
&\text{(using Claim 16)} &\leq\;& \frac{10}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta',\epsilon^{rep}}\cap M\\ h\in T_{m^{t+1}}\\ m^t\in\psi(m^{t+1})}} q_t(m^t)q_t(h|m^t)^2\\
&\text{(assumption in the claim)} &\leq\;& \frac{10}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta',\epsilon^{rep}}\cap M\\ h\in T_{m^{t+1}}\\ m^t\in\psi(m^{t+1})}} q_t(m^t)(a\cdot cer^t(M_t))^2\\
&(|T_{m^{t+1}}|\leq\alpha'|\mathcal{H}|) &\leq\;& \frac{10}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta',\epsilon^{rep}}\cap M\\ m^t\in\psi(m^{t+1})}} q_t(m^t)(a\cdot cer^t(M_t))^2\cdot\alpha'|\mathcal{H}|\\
&(cer^t(M_t)\leq\frac{c}{|\mathcal{H}|}) &\leq\;& 10\alpha' ca^2 cer^t(M_t)\cdot\frac{1}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta',\epsilon^{rep}}\cap M\\ m^t\in\psi(m^{t+1})}} q_t(m^t)\\
&\text{(see below)} &\leq\;& 10\alpha' ca^2 cer^t(M_t)
\end{aligned}$$

to understand why the second inequality is true, notice that we have a sum of the form $4(1+2\epsilon^{rep})\sum_{h\in T} q_{t+1}(h)a_h$ for some value $a_h\geq 0$, which is equal (by the definition of q_t)

to

$$\begin{aligned}
\frac{4(1+2\epsilon^{rep})}{|H_{t+1}|}\sum_{h\in H_{t+1}\cap T} a_h &\leq \frac{4(1+2\epsilon^{rep})}{|H_t|(1-1/c)}\sum_{h\in H_{t+1}\cap T} a_h \\
(\text{for } c\geq 45, \epsilon^{rep}\leq 1/2) &\leq \frac{10}{|H_t|}\sum_{h\in H_{t+1}\cap T} a_h \\
(H_{t+1}\subseteq H_t) &\leq \frac{10}{|H_t|}\sum_{h\in H_t\cap T} a_h \\
&= 10\sum_{h\in T} q_t(h)a_h
\end{aligned}$$

The last inequality is true since every $m \in M$ is in $\psi(m^{t+1})$ for at most $|\mathcal{X}|$ memories m^{t+1}. □

7 Combining Many Sourced and Heavy Sourced Memories

In this section we sum up all the claims proven so far and show that for an hypotheses graph that is d-mixing, if the memory is bounded, then the number of labeled examples used till learning must be large. To do so, we will notice that $cer^0(M_0) = \frac{1}{|\mathcal{H}|}$, and then prove that

$$cer^{t+1}(M_{t+1}) \leq cer^t(M_t)(1+|\mathcal{H}|^{-\nu}),$$

for some small constant $\nu > 0$. This will imply that even after many steps (about $\Omega(|\mathcal{H}|^{\nu})$) the certainty will be at most $c/|\mathcal{H}|$ at each step.

To bound the certainty at each step, we show how to decompose the edges of the knowledge graph, so that each edge leads either to a heavy-sourced memory or to a many-sourced memory (recall Definitions 34, 37), or is part of a *small* error set. To achieve this we duplicate some of the memories. You can find the proof of the next claim in [5].

Claim 44 (Decomposition lemma). *Suppose that the hypotheses graph is an $(\alpha,\beta,\epsilon)-$ H-expander, the number of memory states is at most Λ, and fraction of edges removed from the knowledge graph G_t, i.e., $\gamma = 1-\frac{|E'_t|}{|E_t|}$, is at most 0.5, then for any time t and $\gamma_1,\gamma_2 \in (0,1)$ by*

- *removing at most*

$$\frac{2}{c}+4\beta+4c\gamma_1\gamma_2\Lambda$$

fraction of the edges from G_{t+1} (recall that $c>1$ was used to define Bad_M)

- *creating for each memory m in G_{t+1} copies (m,i) so each edge (m,h) now corresponds to an edge $((m,i),h)$ for some single i*

we can make sure that memories in the new graph G'_{t+1} are only in $M_{t+1}^{many>\gamma_1,\epsilon^{rep}} \cup M_{t+1}^{heavy>\gamma_2}$.

Recall the connection between q_t and G'_t mentioned in Section 3.3 — the probability $q_t(m)$ is the fraction of edges connected to m in G'_t. Notice that in order for this claim to be meaningful, the term $4c\gamma_1\gamma_2\Lambda$ must be smaller than 1.

For all $t \geq 1$, we will construct M_{t+1} formally in the proof of Claim 45. Recall also that H_{t+1} and c were defined in Section 3.5. It might be helpful to think of d in the following claim as roughly $\sqrt{|\mathcal{H}|}$, $c \sim \frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}$, and $|\mathcal{H}| \approx |\mathcal{X}|$.

Claim 45. *For any $\gamma_0 \in (0,1)$, $c > 10^8$, and $k = \log c - 4$, if the hypotheses graph is* d*-mixing, Λ is the number of memory states with $\Lambda \leq \gamma_0^{-k} c^{-7}$, $\frac{c^{102}\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|} \leq 1$, and $2^{k+2}\sqrt{\gamma_0} + 2^{k+2} \cdot c \cdot \sqrt{\frac{16}{\gamma_0^{11}} \frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}} \leq \frac{1}{40c^6}$ then for any time step $t \leq 10^{-8} \cdot c$, the following hold*

- $|H_t| \geq (1 - 1/c)^{t-1}|\mathcal{H}|$
- *the graph G'_t is a $\left(2^{k+2}\sqrt{\gamma_0} + 2^{k+2} \cdot c \cdot \sqrt{\frac{16}{\gamma_0^{11}} \frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}}, \gamma_0^k, (1 + 10\sqrt{\gamma_0} + \frac{1}{c^{15}})(t+1)2^k, \gamma_0, k+1\right)$–K-expander.*
- *for any weight vector w (i.e., $\forall i, w_i \in [0,1]$) on the memories at time t and for any subset of memories at time t, $M \subseteq M_t$*

$$cer_w^t(M) \leq \left[\frac{2^{k+2}}{|\mathcal{H}|}\left(\sum_{m \in M} q_t(m) w_m\right) + \frac{8}{c} \cdot \sum_{t'=1}^{t-1} cer^{t'}(M_{t'})\right]\left(1 + \frac{6}{c}\right)^{t-1}$$

- $q_t(M_t) \geq 1 - \frac{2t}{c}$
- *for each $m^t \in M_t$ it holds that $\sum_h q_t^2(h|m) \leq c \cdot cer^t(M_t)$*
- *for each $h \in H_t, m \in M_t$ it holds that $q_t(h|m) \leq 2c^2 \cdot cer^t(M_t)$*
- *we remove at most $\frac{4t}{c}$ fraction of the edges of the knowledge graph at time t*

Before we prove the claim let us prove (in Claim 46) that the last item in the claim's list implies that $cer^t(M_t) \leq \frac{2^{k+4}}{|\mathcal{H}|} \leq \frac{c}{|\mathcal{H}|}$.

Claim 46. *If for any $t \leq 10^{-8} \cdot c$,*

$$cer^t(M_t) \leq \left[\frac{2^{k+2}}{|\mathcal{H}|} + \frac{8}{c} \cdot \sum_{t'=1}^{t-1} cer^{t'}(M_{t'})\right]\left(1 + \frac{6}{c}\right)^{t-1}$$

then $cer^t(M_t) \leq \frac{2^{k+4}}{|\mathcal{H}|}$.

Proof. First recall a well known inequality, for any x, $1+x \leq e^x \Rightarrow \forall n > 0, (1+x)^n \leq e^{xn}$. Thus, $(1+\frac{6}{c})^t \leq e^{6t/c}$. Since $t \leq 0.001 \cdot c \leq (\ln(2.4/2.3)/6) \cdot c$, we have that $\left(1+\frac{6}{c}\right)^t \leq \frac{2.4}{2.3}$. Thus,

$$cer^t(M_t) \leq \frac{2^{k+3}}{|\mathcal{H}|} + \frac{8.5}{c} \cdot \sum_{t'=1}^{t-1} cer^{t'}(M_{t'}).$$

Let us focus on the following recursively defined series: $a_1 = \frac{2^{k+3}}{|\mathcal{H}|}$ and

$$a_{t+1} = \frac{2^{k+3}}{|\mathcal{H}|} + \frac{8.5}{c} \cdot \sum_{t'=1}^{t} a_{t'}.$$

Then $a_t \geq cer^t(M_t)$. Since this series is monotonically increasing, we have the following upper bound

$$\begin{aligned} a_{t+1} &\leq a_1 + \frac{8.5t}{c} a_t \\ (t \leq 10^{-8} \cdot c) &\leq a_1 + \frac{1}{100} a_t \\ &\leq a_1 + \frac{1}{100}(a_1 + \frac{1}{100} a_{t-1}) \\ (\text{geometric series}) &\leq \ldots \leq 1.02 a_1 \leq \frac{2^{k+4}}{|\mathcal{H}|} \end{aligned}$$

□

Proof. (of Claim 45) From Proposition 10 we know that the hypotheses graph is an $(\epsilon_{sam}, \epsilon'_{sam} = \frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|\epsilon_{sam}^2})$-sampler for any $\epsilon_{sam} > 0$. From Proposition 12, it is also $(\alpha, \beta, \epsilon) -$ H-expander with $\beta = \frac{c^{100}\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}$ and $\epsilon = \frac{2\mathrm{d}}{\sqrt{\alpha|\mathcal{H}|\beta|\mathcal{X}|}}$ for any α. We pick $\alpha = 1/c^{34}$. By the choice of α, β and for $c \geq 2$ we have that

$$\epsilon = \frac{2\mathrm{d}}{\sqrt{\alpha|\mathcal{H}| \cdot \frac{c^{100}\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|} \cdot |\mathcal{X}|}} = \frac{2}{\sqrt{\alpha c^{100}}} \leq \frac{1}{c^{17}}.$$

Note that since $k = \log c - 4$, the certainty is bounded by $c/|\mathcal{H}|$ (see Claim 46). Denote

$$\epsilon^{rep} := 4\sqrt{\alpha|\mathcal{H}|c \cdot cer^t(M_t)} + 4\epsilon \leq 4c\sqrt{\alpha} + 4\epsilon \leq \frac{4}{c^{16}} + \frac{4}{c^{17}} \leq \frac{1}{c^{15}}.$$

We prove the claim by induction on t.

Induction Basis. At the beginning , before the algorithm got an example, $H_0 = \mathcal{H}$, the certainty of each memory m is $cer^0(m) = \frac{1}{|\mathcal{H}|}$, M_0 contains all the memories, and G'_0 is a $(\alpha', \beta', 2^{\log_{\gamma_0}\beta'}, \gamma_0, k+1) -$ K-expander for any $\alpha', \beta', \gamma_0 \in (0,1)$ and k.

Induction Step. We use the known inequality $1 - x \geq e^{-2x}$ for $x \in (0, 1/2) \Rightarrow \forall n > 0, (1-x)^n \geq e^{-2xn}, x \in (0, 1/2)$, and Claim 18 to deduce that (recall $c \geq 2$)

$$|H_t| \geq (1 - 1/c)^{t-1}|\mathcal{H}| \geq e^{-2(t-1)/c}|\mathcal{H}| \geq e^{\ln 1/3}|\mathcal{H}| = \frac{|\mathcal{H}|}{3},$$

where the third inequality holds since $t - 1 \leq 0.5 \cdot c \leq \frac{c \ln 3}{2}$.

Using Claim 28 and the inductive hypothesis, the graph G'_{t+1} is a

$$\left(2^{k+2}\sqrt{\gamma_0} + c \cdot \sqrt{\frac{16}{\gamma_0^{11}} \frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}}, \gamma_0^k, (1 + 10\sqrt{\gamma_0} + 2\epsilon^{rep})(t+2)2^k, \gamma_0, k+1\right) - \text{K-expander.}$$

From the the inductive hypothesis we have that at most a fraction of $\frac{4t}{c} \leq 0.5$ edges were removed from G'_t.

We use Claim 44 with

- Let γ_1 define the many-source set $M_{t+1}^{many>\gamma_1,\epsilon^{rep}}$ (see Definition 37). To later apply Claim 43, we choose $\gamma_1 = \gamma_0^k$.
- Let γ_2 define the heavy-source set $M_{t+1}^{heavy>\gamma_2}$ (see Definition 34). To later apply Claim 36 we choose
$$\gamma_2 = 5\epsilon_{sam}c + 20c^6\sqrt{\frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|\epsilon_{sam}^2}}$$
We choose ϵ_{sam} such that γ_2 will be minimized. To do so, we equate the two terms that comprise γ_2 by choosing $\epsilon_{sam}^2 = 4c^5\sqrt[2]{\frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}}$, which means that $\gamma_2 < 10c^4\sqrt[4]{\frac{\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}}$.

For later use, notice that

$$\gamma_1\gamma_2 \leq \gamma_0^k 10c^4\sqrt[4]{\frac{\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}}.$$

From Claim 44 we know that by removing at most

$$\frac{2}{c} + 4\beta + \gamma_0^k 40c^5\sqrt[4]{\frac{\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}}\Lambda$$

fraction of the edges, the graph only has heavy-sourced or many-sourced memories.

Fix M a set of memories in G'_{t+1} and a weight vector w (i.e., for each memory at time $t+1$, w assigns a weight in $[0, 1]$)

Heavy-sourced memories. We can use Claim 36 to deduce that

$$\begin{aligned}
cer_w^{t+1}(M_{t+1}^{heavy>\gamma_2}\cap M) &\leq \left[\left(1+\frac{4}{c}\right)\sum_{(m^t,S)\in M_{t+1}^{heavy>\gamma_2}\cap M} cer^t(m^t)\frac{|S|}{|\mathcal{X}|}w_{(m^t,S)}\right]+\left[\frac{2}{c}\cdot cer^t(M_t)\right]\\
&\leq \left[\sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>\gamma_2}\cap M\\ h\in\mathcal{H}}} cer^t(m^t)\frac{|S|}{|\mathcal{X}|}w_{(m^t,S)}\right]+\left[\frac{6}{c}\cdot cer^t(M_t)\right]\\
\text{(see below)} &\leq \left[\sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>\gamma_2}\cap M\\ h\in\mathcal{H}}} q_t(m^t)q_t^2(h|m^t)\left(1+\frac{1}{c}\right)q_{t+1}(S|m^t)w_{(m^t,S)}\right]+\\
&\quad\left[\frac{6}{c}\cdot cer^t(M_t)\right]\\
&\leq \left[\sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>\gamma_2}\cap M\\ h\in\mathcal{H}}} q_t(m^t)q_t^2(h|m^t)q_{t+1}(S|m^t)w_{(m^t,S)}\right]+\left[\frac{7}{c}\cdot cer^t(M_t)\right] \qquad (\star)
\end{aligned}$$

To prove the third inequality we will show that for $|S|\geq\gamma_2|\mathcal{X}|$ it holds that

$$\frac{|S|}{|\mathcal{X}|}\leq\left(1+\frac{1}{c}\right)q_{t+1}(S|m^t).$$

From the sampler property (see Definition 7) we know that for each subset of labeled examples $S\subseteq\mathcal{X}$ there is a set $Err(S)\subseteq\mathcal{H}$ with $|Err(S)|\leq\epsilon'_{sam}|\mathcal{H}|$ such that for each $h\notin Err(S)$,

$$\Pr(S|h)\geq\frac{|S|}{|\mathcal{X}|}-\epsilon_{sam}$$

From Claim 22

$$\begin{aligned}
q_{t+1}(S|m^t) &= \sum_h \Pr(S|h) q_t(h|m^t) \\
&\geq \sum_{h \notin Err(S)} \Pr(S|h) q_t(h|m^t) \\
&\geq \sum_{h \notin Err(S)} \left(\frac{|S|}{|\mathcal{X}|} - \epsilon_{sam} \right) q_t(h|m^t) \\
&= \frac{|S|}{|\mathcal{X}|} \left(1 - \frac{\epsilon_{sam}}{|S|/|\mathcal{X}|} \right) \sum_{h \notin Err(S)} q_t(h|m^t) \\
\text{(definition of } \gamma_2 \text{)} \quad &\geq \frac{|S|}{|\mathcal{X}|} \left(1 - \frac{\epsilon_{sam}}{5\epsilon_{sam} c} \right) \sum_{h \notin Err(S)} q_t(h|m^t) \\
\text{(Claim 2 \&} cer^t(M_t) \leq \frac{c}{|\mathcal{H}|}\text{)} \quad &\geq \frac{|S|}{|\mathcal{X}|} \left(1 - \frac{1}{5c} \right) (1 - c\sqrt{\epsilon'_{sam}})
\end{aligned}$$

This means that

$$\frac{|S|}{|\mathcal{X}|} \leq \frac{q_{t+1}(S|m^t)}{\left(1 - \frac{1}{5c}\right)(1 - c\sqrt{\epsilon'_{sam}})}.$$

So we just need to show that

$$\frac{1}{\left(1 - \frac{1}{5c}\right)(1 - c\sqrt{\epsilon'_{sam}})} \leq 1 + \frac{1}{c}$$

Note that $c\sqrt{\epsilon'_{sam}} \leq 1/4c$ since

$$\epsilon'_{sam} = \frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|\epsilon_{sam}^2} = \frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}| 4c^5 \sqrt[2]{\frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}}} = \frac{1}{4c^5} \sqrt[2]{\frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}} \leq \frac{1}{16c^4}$$

Also note that

$$\begin{aligned}
\frac{1}{\left(1 - \frac{1}{5c}\right)\left(1 - \frac{1}{4c}\right)} - 1 &= \frac{1 - (1 - 1/(4c) - 1/(5c) + 1/(20c^2))}{\left(1 - \frac{1}{5c}\right)\left(1 - \frac{1}{4c}\right)} \\
\left(\frac{1}{(1 - \frac{1}{5c})(1 - \frac{1}{4c})} \leq 2 \right) \quad &\leq 2(1/(4c) + 1/(5c)) \\
&\leq \frac{1}{c}
\end{aligned}$$

Many-sourced memories. We can use Claim 43 and get that

$$\begin{aligned} cer^{t+1}(M_{t+1}^{many>\gamma_1,\epsilon^{rep}} \cap M) &\leq \frac{2^{k+2}}{|\mathcal{H}|} \cdot \sum_{m \in M_{t+1}^{many>\gamma_1,\epsilon^{rep}} \cap M} q_{t+1}(m) w_m + \frac{1}{c} \cdot cer^t(M_t) \\ &= \left[\frac{2^{k+2}}{|\mathcal{H}|} \cdot \sum_{\substack{m=\{(m_i^t,x_i)\} \in \\ M_{t+1}^{many>\gamma_1,\epsilon^{rep}} \cap M}} q_t(m_i^t) q_{t+1}(x_i|m_i^t) w_m \right] + \frac{1}{c} \cdot cer^t(M_t) \qquad (\star\star) \end{aligned}$$

Combining heavy-sourced and many-sourced memories. For each m^t, memory at time t, we define the weight of m^t due to heavy-sourced memories

$$w_{m^t}^{heavy} := \sum_{S|(m^t,S) \in M_{t+1}^{heavy>\gamma_2} \cap M} q_{t+1}(S|m^t) w_{(m^t,S)}.$$

Similarly, we define the weight of m^t due to many-sourced memories

$$w_{m^t}^{many} := \sum_{x_i|m=\{(m^t,x_i)\} \in M_{t+1}^{many>\gamma_1,\epsilon^{rep}} \cap M} q_{t+1}(x_i|m^t) w_m.$$

The total weight of m^t is denoted by $w_{m^t} = w_{m^t}^{heavy} + w_{m^t}^{many}$. Combining $(\star), (\star\star)$ we have that

$$\begin{aligned} cer_w^t(M) &\leq \sum_{m^t} q_t(m^t) \left(\sum_h q_t^2(h|m^t) \cdot w_{m^t}^{heavy} + \frac{2^{k+2}}{|\mathcal{H}|} \cdot w_{m^t}^{many} \right) + \frac{8}{c} \cdot cer^t(M_t) \\ &\leq \sum_{m^t} q_t(m^t) \max \left\{ \sum_h q_t^2(h|m^t), \frac{2^{k+2}}{|\mathcal{H}|} \right\} \cdot w_{m^t} + \frac{8}{c} \cdot cer^t(M_t) \end{aligned}$$

Define $M_a = \{m^t | \sum_h q_t^2(h|m^t) > 2^{k+2}/|\mathcal{H}|\}$ and $M_b = \{m^t | \sum_h q_t^2(h|m^t) \leq 2^{k+2}/|\mathcal{H}|\}$ and the last term is equal to

$$\left[\sum_{m^t \in M_a} q_t(m^t) w_{m^t} \cdot \sum_h q_t^2(h|m^t) \right] + \left[\sum_{m^t \in M_b} q_t(m^t) w_{m^t} \cdot \frac{2^{k+2}}{|\mathcal{H}|} \right] + \left[\frac{8}{c} \cdot cer^t(M_t) \right]$$

using the induction hypothesis on M_a, the last expression is at most

$$\left[\frac{2^{k+2}}{|\mathcal{H}|} \left(\sum_{m^t \in M_a} q_t(m^t) w_{m^t} \right) + \frac{8}{c} \cdot \sum_{t'=1}^{t-1} cer^{t'}(M_{t'}) \right] \left(1 + \frac{6}{c}\right)^{t-1} + \left[\sum_{m^t \in M_b} q_t(m^t) w_{m^t} \cdot \frac{2^{k+2}}{|\mathcal{H}|} \right] + \left[\frac{8}{c} \cdot cer^t(M_t) \right]$$

which is at most

$$\left[\frac{2^{k+2}}{|\mathcal{H}|}\left(\sum_{m^t\in M_t} q_t(m^t)w_{m^t}\right)+\frac{8}{c}\cdot\sum_{t'=1}^{t} cer^{t'}(M_{t'})\right]\left(1+\frac{6}{c}\right)^{t-1}$$

and we get the bound we wanted to show using the following equalities

$$\begin{aligned}
\sum_{m\in M} q_{t+1}(m)w_m &= \sum_{(m^t,S)\in M_{t+1}^{heavy>\gamma_2}\cap M} q_t(m^t)q_{t+1}(S|m^t)w_{(m^t,S)} + \\
&\quad \sum_{m=\{(m_i^t,x_i)\}\in M_{t+1}^{many>\gamma_1,\epsilon^{rep}}\cap M} q_t(m_i^t)q_{t+1}(x_i|m_i^t)w_m \\
&= \sum_{m^t} q_t(m^t) \sum_{S|(m^t,S)\in M_{t+1}^{heavy>\gamma_2}\cap M} q_{t+1}(S|m^t)w_{(m^t,S)} + \\
&\quad \sum_{m^t} q_t(m^t) \sum_{x_i|m=\{(m_i^t,x_i)\}\in M_{t+1}^{many>\gamma_1,\epsilon^{rep}}\cap M} q_{t+1}(x_i|m^t)w_m \\
&= \sum_{m^t\in M_t} q_t(m^t)w_{m^t}^{heavy} + \sum_{m^t\in M_t} q_t(m^t)w_{m^t}^{many} \\
&= \sum_{m^t\in M_t} q_t(m^t)w_{m^t}
\end{aligned}$$

Removing edges. Denote by M' all memories at time $t+1$ that are heavy-sourced or many-sourced. So far we bounded the average certainty $cer^{t+1}(M')$. Notice that this average certainty is equal to

$$cer^{t+1}(M') = \sum_{m\in M',h\in\mathcal{H}} q_{t+1}(m,h)q_{t+1}(h|m).$$

Applying Markov's inequality, we have that

$$\Pr_{h,m}[q_{t+1}(h|m)\geq c^2\cdot cer^{t+1}(M')]\leq\frac{1}{c^2}.$$

We will remove all edges with $q_{t+1}(h|m)\geq c^2\cdot cer^{t+1}(M')$. We will show that this removal does not increase the certainty by much for most memories.

Denote by Err all pairs (m,h) such that $q_{t+1}(h|m)\geq c^2\cdot cer^{t+1}(M)$. Putting in different words the last equation, we have that

$$\sum_{m} q_{t+1}(m)\left[\sum_{h|(m,h)\in Err} q_{t+1}(h|m)\right]\leq\frac{1}{c^2}.$$

Applying Markov's inequality again, we have that for most memories we do not delete too many edges:

$$\Pr_m\left[\sum_{h|(m,h)\in Err} q_{t+1}(h|m) > \frac{1}{c}\right] \leq \frac{1}{c}$$

As was promised in Section 3.5, we maintain a substantial set of memories $M_{t+1} \subseteq \mathcal{M}$ that we focus on, and we are ready to define it

$$M_{t+1} := \left\{ m \in M' \;\middle|\; \sum_{h|(m,h)\in Err} q_{t+1}(h|m) \leq \frac{1}{c} \text{ and } \sum_h q_t^2(h|m) \leq c \cdot cer^t(M_t) \right\},$$

recall that M' contains all the memories that are heavy-sourced or many-sourced. Thus, using also Claim 15, we have that

$$q_{t+1}(M_{t+1}) \geq q_t(M_t) - \frac{2}{c} \geq 1 - \frac{2(t+1)}{c}.$$

Note that for all $m \in M_{t+1}$, the removal of edges with $q_{t+1}(h|m) \geq c^2 \cdot cer^{t+1}(M')$ can only increase by at most a factor of $\frac{1}{1-1/c} \leq 1 + \frac{1.1}{c}$ the probability $q_{t+1}(h|m)$ (because we have removed at most $1/c$ fraction of the edges from $m \in M_{t+1}$). Thus, for each $m \in M_{t+1}$ $q_{t+1}(h|m) \leq \left(1 + \frac{1.1}{c}\right) c^2 cer^{t+1}(M_{t+1}) \leq 2c^2 cer^{t+1}(M_{t+1})$.

Let us now also remove the edges from Claim 44. Thus (using the bound we showed earlier on $\gamma_1\gamma_2$), in time $t+1$ we removed a total fraction of

$$\left(\frac{1}{c} + \frac{1}{c^2}\right) + \left(\frac{2}{c} + 4\beta + \gamma_0^k 40c^5 \sqrt[4]{\frac{\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}}\Lambda\right) \leq \frac{1}{c} + \frac{1}{16c} + \frac{2}{c} + \frac{1}{4c} + \frac{1}{2c} \leq \frac{4}{c}.$$

edges.

The last removal increases the average certainty $cer_w^{t+1}(M)$ by at most $(1 + 4/c)$. So, in total, the removals cause the average certainty $cer_w^{t+1}(M)$ to increase by a factor of at most $(1 + 4/c) \cdot (1 + 1.1/c) \leq 1 + \frac{6}{c}$. To sum up,

$$cer_w^{t+1}(M) \leq \left[\frac{2^{k+2}}{|\mathcal{H}|}\left(\sum_{m\in M} q_t(m) w_m\right) + \frac{8}{c} \cdot \sum_{t'=1}^{t} cer^{t'}(M_{t'})\right]\left(1 + \frac{6}{c}\right)^t$$

□

7.1 Choosing Parameters

Before we prove the main theorem, in Claim 47 we prove that $\mathrm{d} = \Omega(\sqrt{|\mathcal{X}|})$. The proof is a straightforward adaptation of a result in [2]. Let $S \subseteq \mathcal{X}$ and $T \subseteq \mathcal{H}$. Define

$$dis(S,T) := |E(S,T)| - \frac{|S||T|}{2}$$

Claim 47. *For every $0 < \epsilon < 1/16$ there exists $\epsilon' = 10^{-2}\epsilon^{3/2}$ such that there are $S \subseteq \mathcal{X}, T \subseteq \mathcal{H}$ with $|S| \leq \epsilon|\mathcal{X}|, |T| \leq \epsilon|\mathcal{H}|$ and*

$$|dis(S,T)| > \epsilon' \sqrt{|\mathcal{X}|}\sqrt{|\mathcal{H}||\mathcal{X}|},$$

which implies that if the hypotheses graph is d*-mixing then* $\mathrm{d} \geq \frac{\epsilon'}{\epsilon}\sqrt{|\mathcal{X}|} = 10^{-2}\sqrt{\epsilon} \cdot \sqrt{|\mathcal{X}|}$.

Proof. Let $T \subseteq \mathcal{H}$ with $|T| = \epsilon|\mathcal{H}|$ be a randomly chosen and let $x \in \mathcal{X}$ be a random labeled example. Then,

$$\Pr_{T,v}\left[\left||\Gamma(x) \cap T| - \frac{|\mathcal{X}|}{2}\right| > 10^{-2}\sqrt{\epsilon\frac{|\mathcal{H}|}{2}}\right] > \frac{1}{2}.$$

Define

$$V(T) = \left\{x \in \mathcal{X} : \left||\Gamma(x) \cap T| - \frac{|\mathcal{X}|}{2}\right| > 10^{-2}\sqrt{\epsilon\frac{|\mathcal{H}|}{2}}\right\} \subseteq \mathcal{X}.$$

The expected size of $V(T)$ equals to

$$\sum_{x \in \mathcal{X}} \Pr_{T,v}\left[\left||\Gamma(x) \cap T| - \frac{|\mathcal{X}|}{2}\right| > 10^{-2}\sqrt{\epsilon\frac{|\mathcal{H}|}{2}}\right] > \frac{|\mathcal{X}|}{2}.$$

Hence,

$$\frac{|\mathcal{X}|}{2} < \mathbb{E}[|V(T)|] \leq |\mathcal{X}| \Pr\left[|V(T)| > \frac{|\mathcal{X}|}{4}\right] + \frac{|\mathcal{X}|}{4}\left(1 - \Pr\left[|V(T)| > \frac{|\mathcal{X}|}{4}\right]\right),$$

implying

$$\Pr\left[|V(T)| > \frac{|\mathcal{X}|}{4}\right] > \frac{1}{3}.$$

Thus, one can choose a specific T and $S \subseteq V(T)$ with $|S| = \epsilon|\mathcal{X}|$ such that $dis(x,T) > 10^{-2}\sqrt{\epsilon\frac{|\mathcal{H}|}{2}}$ or $dis(x,T) < -10^{-2}\sqrt{\epsilon\frac{|\mathcal{H}|}{2}}$ hold for all $x \in S$. In both cases S and T have $|dis(S,T)| > 10^{-2}\epsilon^{3/2}\sqrt{|\mathcal{X}|} \cdot \sqrt{|\mathcal{H}||\mathcal{X}|}$. □

For a mixing hypothesis class $\mathcal{H}$, i.e., when $d^2 \approx |\mathcal{X}|$, we show a lower bound of $\Omega(\log^2|\mathcal{H}|)$ on the number of bits needed for learning $\mathcal{H}$ using less than $|\mathcal{H}|^{\Theta(1)}$ labeled examples.

Theorem 48 (main theorem)**.** *If the hypotheses graph is* d*-mixing,* $m := \frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}$ *and* $|\mathcal{H}|$ *are at least some constants, then any learning algorithm that outputs the underlying hypothesis with probability at least* $m^{-\Theta(1)}$ *must use at least* $2^{\Omega(\log^2 m)}$ *memory states or* $m^{\Omega(1)}$ *labeled examples.*

Proof. To apply Claim 45 to $\mathcal{H}$ we will make sure that the following hold:

1. $\frac{c^{102}\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|} \leq 1$ by choosing $c = \sqrt[207]{\frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}}$ (note that by definition of mixing $\mathrm{d} \leq \sqrt{|\mathcal{H}||\mathcal{X}|}$, hence $c > 1$)

2. $2^{k+2}\sqrt{\gamma_0} + 2^{k+2} \cdot c \cdot \sqrt{\frac{16}{\gamma_0^{11}} \frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}} \leq \frac{1}{40c^6}$ with $k = \log c - 12$: we will show that

   (a) $2^{k+2}\sqrt{\gamma_0} \leq \frac{1}{80c^6} \Leftrightarrow 2^{-10}c^7\sqrt{\gamma_0} \leq 1 \Leftarrow \gamma_0 \leq \frac{1}{c^{14}}$

   (b) $2^{k+2} \cdot c \cdot \sqrt{\frac{16}{\gamma_0^{11}} \frac{\mathrm{d}^2}{|\mathcal{X}||\mathcal{H}|}} \leq \frac{1}{80c^6} \Leftrightarrow \frac{c^2}{2^{10}} \cdot \sqrt{\frac{16}{\gamma_0^{11}} \frac{1}{c^{207}}} \leq \frac{1}{80c^6} \Leftarrow \frac{16 \cdot 80^2}{2^{20}} \frac{1}{c^{207-16}} \leq \gamma_0^{11} \Leftarrow \frac{1}{c^{17}} \leq \gamma_0$

   by choosing $\gamma_0 = 1/c^{15}$.

3. $c > 10^8 \Leftarrow \frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2} \geq 10^{1700}$

4. $\Lambda \leq \gamma_0^{-k} c^{-7} = c^{15k-7} = c^{15 \log c - 187}$

From Claim 45 we can deduce that even after $t \leq 10^{-8} \cdot c$ examples given, the certainty is at most $c/|\mathcal{H}|$. We pick $t := \sqrt{c} \leq 10^{-8} \cdot c$ (for large enough c). The total number of edges removed is at most $\frac{4t}{c}$, and $1 - q_t(M_t) \leq \frac{2t}{c}$. Using Claim 19 there is an hypothesis $h \in \mathcal{H}$ such that the probability to correctly return it is at most

$$3\sqrt{c \cdot \frac{c}{|\mathcal{H}|}} + 3 \cdot \frac{2t}{c} + \frac{4t}{c}$$

Let us bound this expression. We will bound the first term by $3/c$ by proving that $c^4 \leq |\mathcal{H}|$. Since $\mathrm{d}^2 \geq 10^{-6}|\mathcal{X}|$ (see Claim 47), $c^4 \leq (|\mathcal{H}|10^6)^{4/207}$ which is smaller than $|\mathcal{H}|$. The sum of the last two terms is $6t/c = 6/\sqrt{c}$.

□

The next theorem proves that even if we allow the bounded-algorithm to return an approximation of the underlying hypothesis it still needs many examples.

Theorem 49. *If the hypotheses graph is* d*-mixing,* $m := \frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}$ *and* $|\mathcal{H}|$ *are at least some constants, then any learning algorithm that outputs the underlying hypothesis, or an approximation of it, with probability at least* $m^{-\Theta(1)}$ *must use at least* $2^{\Omega(\log^2 m)}$ *memory states or* $m^{\Omega(1)}$ *labeled examples.*

Proof. From Claim 21 in [5], we know that there is an hypothesis class $\mathcal{H}' \subseteq \mathcal{H}$ with $|\mathcal{H}'| \geq \frac{|\mathcal{H}|}{1 + \frac{16\mathrm{d}^2}{|\mathcal{X}|}} \geq \frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2 10^6}$ (for the last inequlity see Claim 47) such that every two hypotheses in $\mathcal{H}'$ has agreement less than $3/4$.

We apply Theorem 48 to $\mathcal{H}'$. Thus, $\mathcal{H}'$ is unlearnable with bounded memory (since all hypotheses in $\mathcal{H}'$ are far apart). Note that the learner is even unable to improper learn

$\mathcal{H}'$ (which means that the learner can return hypothesis not in $\mathcal{H}'$) — because the learner does not have any computational limitations, it can compute an hypothesis in $\mathcal{H}'$ exactly (since all hypotheses in $\mathcal{H}'$ are far apart). This implies that also $\mathcal{H}$ is unlearnable with bounded memory.

□

References

[1] B. Chazelle. *The Discrepancy Method: Randomness and Complexity.* Randomness and Complexity. Cambridge University Press, 2000.

[2] P. Erdós, M. Goldberg, J. Pach, and J. Spencer. Cutting a graph into two dissimilar halves. *Journal of graph theory*, 12(1):121–131, 1988.

[3] G. Kol, R. Raz, and A. Tal. Time-space hardness of learning sparse parities. In *Proc. 49th ACM Symp. on Theory of Computing*, 2017.

[4] M. Krivelevich and B. Sudakov. Pseudo-random graphs. In *More sets, graphs and numbers*, pages 199–262. Springer, 2006.

[5] D. Moshkovitz and M. Moshkovitz. Mixing implies lower bounds for space bounded learning. Technical report, ECCC Report TR17-017, 2017.

[6] 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.

[7] R. Raz. A time-space lower bound for a large class of learning problems. Technical report, ECCC Report TR17-020, 2017.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il