Proceedings of Machine Learning Research vol 65:1–51, 2017

Mixing Implies Lower Bounds for Space Bounded Learning

Dana Moshkovitz DANAMA@CS.UTEXAS.EDU
Department of Computer Science, UT Austin *

Michal Moshkovitz MICHAL.MOSHKOVITZ@MAIL.HUJI.AC.IL
The Edmond and Lily Safra Center for Brain Sciences, Hebrew University †

Abstract

One can learn any hypothesis class $\mathcal{H}$ with $O(\log|\mathcal{H}|)$ labeled examples. Alas, learning with so few examples requires saving the examples in memory, and this requires $|\mathcal{X}|^{O(\log|\mathcal{H}|)}$ memory states, where $\mathcal{X}$ is the set of all labeled examples. This motivates the question of how many labeled examples are needed in case the memory is bounded.

Previous work showed, using techniques such as linear algebra and Fourier analysis, that parities cannot be learned with bounded memory and less than $|\mathcal{H}|^{\Omega(1)}$ examples. One might wonder whether a general combinatorial condition exists for unlearnability with bounded memory, as we have with the condition $VCdim(\mathcal{H}) = \infty$ for PAC unlearnability.

In this paper we give such a condition. We show that if an hypothesis class $\mathcal{H}$, when viewed as a bipartite graph between hypotheses $\mathcal{H}$ and labeled examples $\mathcal{X}$, is mixing, then learning it requires $|\mathcal{H}|^{\Omega(1)}$ examples under a certain bound on the memory. Note that the class of parities is mixing. Moreover, as an immediate corollary, we get that most hypothesis classes are unlearnable with bounded memory. Our proof technique is combinatorial in nature and very different from previous analyses.

Keywords: Bounded space, Lower bound, Mixing, PAC learning, Time-space tradeoff, VC-dimension

1. Introduction

1.1. Space Bounded Learning

In learning theory one wishes to learn an hypothesis drawn from a class $\mathcal{H}$ of hypotheses by receiving random labeled examples (Valiant, 1984). For simplicity, we focus on finite classes of Boolean hypotheses. For instance, $\mathcal{H}$ can be the family of parities $\oplus_{i\in S}x_i$ for different $S \subseteq \{1,\ldots,n\}$, and random labeled examples are pairs $(\langle x_1,\ldots,x_n\rangle, \oplus_{i\in S}x_i)$ for random $\langle x_1,\ldots,x_n\rangle \in \{0,1\}^n$. We denote by $\mathcal{X}$ the family of labeled examples, so $|\mathcal{H}| = 2^n$ and $|\mathcal{X}| = 2^{n+1}$.

The fundamental theorem of statistical learning implies that learning is possible after seeing $O(\log|\mathcal{H}|)$ labeled examples, since most labeled examples would cut the number of feasible hypotheses by a factor of about two. Alas, learning with so few examples requires saving the examples in memory, and this requires $|\mathcal{X}|^{O(\log|\mathcal{H}|)}$ memory states. A recent line of work asks how the number of required labeled examples changes if we restrict the memory of the learner (see Shamir (2014); Steinhardt et al. (2016)).

* This material is based upon work supported by the National Science Foundation under grants number 1218547 and 1648712.

† This work is partially supported by the Gatsby Charitable Foundation, The Israel Science Foundation, and The Harry and Sylvia Hoffman Leadership and Responsibility Program.

To understand the quantitative aspects of space bounded learning, we'll need two observations:

- *Memory states:* At least $|\mathcal{H}|$ memory states are needed in order to exactly distinguish $|\mathcal{H}|$ possible hypotheses. The focus is on bounds on the number of memory states that are significantly larger than $|\mathcal{H}|$.
- *Examples:* At most $O(|\mathcal{H}| \log |\mathcal{H}|)$ examples always suffice for learning, even if the number of memory states is only $|\mathcal{H}|$, since the learner can enumerate the hypotheses one by one, ruling out the current hypothesis if an inconsistent labeled example comes up. The question is whether one can prove a lower bound on the number of examples that comes close to $\approx |\mathcal{H}|$.

We refer to hypothesis classes that require $|\mathcal{H}|^{\Omega(1)}$ examples for learning with bounded space as *unlearnable* under the space constraint. Note that $|\mathcal{H}|$ is exponentially worse than the number of examples $O(\log |\mathcal{H}|)$ sufficient without memory constraints.

Raz (2016) was the first to prove lower bounds on space-bounded learning. He considered the class of parities mentioned above, as suggested in Steinhardt et al. (2016). Raz showed that either $|\mathcal{X}|^{\Omega(\log|\mathcal{H}|)} = 2^{\Omega(n^2)}$ memory states or $|\mathcal{H}|^{\Omega(1)} = 2^{\Omega(n)}$ examples are needed for learning this class. In other words, he showed that parities are unlearnable with $2^{o(n^2)}$ memory states. His work was then generalized in Kol et al. (2017) to parities on $l \leq n/2$ variables, and this in turn gives lower bounds for classes that contain such parities. Raz's work and its generalization were constrained to learning parities and used techniques such as linear algebra and Fourier analysis. This begs the question of proving lower bounds for more general hypothesis classes.

1.2. This Work

We give a simple, combinatorial, sufficient condition for a Boolean hypothesis class to be unlearnable with sufficiently bounded memory. The condition includes parity classes and is about the "mixing" properties of the hypothesis class when viewed as a graph, as explained next.

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., $h(x) = y$. Mixing is defined as follows, similarly to the Expander Mixing Lemma (Hoory et al., 2006):

Definition 1 (Mixing) *We say that a bipartite graph (A, B, E) with average left degree $\bar{d}_A$ is* d*-mixing if for any $A' \subseteq A, B' \subseteq B$ it holds that*

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

For example, for parities $\mathrm{d} = \sqrt{|\mathcal{H}|}$ and $|\mathcal{H}| = 2|\mathcal{X}|$. In general we say that an hypothesis class is mixing if the corresponding bipartite graph is $O(\sqrt{|\mathcal{X}|})$-mixing. Mixing classes are such that, even if one knows that the underlying hypothesis h was taken from a (sufficiently large) subset A' of the hypotheses, knowing that h is consistent with at least one example of a (sufficiently large) set B' of labeled examples reveals very little further information on h besides its membership in A'. Additionally, for mixing classes, even an approximation of the underlying hypothesis typically uniquely determines the hypothesis, since (for the most part) different hypotheses differ on a substantial number of labeled examples (see the paper for a formal statement and proof). Hence, PAC-learning implies exact learning for mixing classes.

A learning algorithm that has Λ memory states and uses T labeled examples is a branching program of length T and width Λ, given by a directed multi-graph with vertices in $T+1$ layers containing Λ vertices each. The algorithm starts with an arbitrary vertex in the first layer. Each vertex, except those in the last layer, has out degree exactly $|\mathcal{X}|$ and is marked with a unique labeled example. The algorithm transitions from one memory state to another according to the labeled example it received. When the learning algorithm reaches the last layer it outputs an hypothesis that depends on the memory state it ended up with.

We prove that mixing hypothesis classes are unlearnable when the memory is bounded by roughly $|\mathcal{H}|^{1.25}$ memory states.

Theorem 2 (Main theorem) *For any constant $s_1 \in (0,1)$, there are constants $s_2, s_3 > 0$ such that if the hypotheses graph is d-mixing, $|\mathcal{H}| \geq s_2$, and the number of memory states is bounded by*

$$\left(\frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}\right)^{1.25} \cdot \frac{1}{\left(1+\frac{16\mathrm{d}^2}{|\mathcal{X}|}\right)^{1.25}|\mathcal{H}|^{s_1}},$$

then any learning algorithm that returns the underlying hypothesis with probability at least $1/3$ must observe at least $|\mathcal{H}|^{s_3}$ labeled examples.

In Section 12 we show two consequences of the main theorem. One immediate consequence that uses the fact that random graphs are mixing (see e.g., Krivelevich and Sudakov (2006)), is that almost all hypothesis classes are unlearnable with bounded memory. Note that unlike for circuits, such a result does not follow from counting arguments[1]. Another consequence of the main theorem is that any hypothesis class that forms an "error correcting code" (i.e., any two hypotheses in $\mathcal{H}$ do not agree on many examples) cannot be learned with bounded memory.

Our work provides a general framework for proving lower bounds on space bounded learning. The framework is combinatorial and fundamentally different from Raz's analysis. In the next subsection we compare our results to previous work. We end the introduction with an outline of the proof.

1.3. Previous Work

Shamir (in a private communication) and later, and independently, Steinhardt, Valiant, and Wager Steinhardt et al. (2016) asked if one can show a lower bound on the number of examples needed, given an upper bound on the number of memory states. Specifically, Steinhardt et al. (2016) focused on the class of parities and conjectured that $|\mathcal{H}|^{\Omega(1)}$ examples are needed if the number of memory states is sufficiently smaller than $|\mathcal{X}|^{\log|\mathcal{H}|}$. Raz (2016) proved the conjecture of Steinhardt et al. (2016), thus showing a tight lower bound for parities. Later this work was generalized in Kol et al. (2017), using similar techniques, to parities on $l \leq n/2$ variables. Unlike those previous works that were limited to parities, we provide a general framework to prove lower bounds on space bounded learning that works for all mixing classes. Parities are mixing, as are random hypothesis classes. For the latter previous techniques did not apply. Moreover, our proof is combinatorial and fundamentally different from those in Raz (2016); Kol et al. (2017). Our method suggests

1. The number of possible hypotheses classes is $\binom{2^{|\mathcal{X}|}/2}{|\mathcal{H}|} \leq 2^{|\mathcal{X}||\mathcal{H}|}$, whereas the number of learners with Λ memory states and T labeled examples is about $\Lambda^{\Lambda T|\mathcal{X}|}$. For parameters of interest, like $\Lambda = |\mathcal{H}|^{\Theta(1)}$ and $T = |\mathcal{H}|^{\Theta(1)}$, the number of learners is much larger than the number of hypotheses classes.

a combinatorial sufficient condition for unlearnability with bounded memory, as we have with the condition $VCdim(\mathcal{H}) = \infty$ for PAC unlearnability. The downside of the result in the main theorem is that the bound on the number of memory states is only roughly $|\mathcal{H}|^{1.25}$ as opposed to $|\mathcal{H}|^{\Omega(\log|\mathcal{X}|)}$ of Raz (2016). We hope that by building on the new framework we present it will be possible to prove optimal lower bounds for wide classes of hypotheses.

1.4. Related Work

Raz established a result closely related to ours Raz (2017) shortly after our paper appeared as a technical report on the ECCC. Like our work, Raz shows that a mixing condition for the hypothesis class implies time-space tradeoffs for learning the class. Comparing the two results:

- Combinatorial vs. Algebraic mixing: Our result is premised on a combinatorial mixing condition (on the number of edges between every two large sets of vertices) whereas Raz's result is based on a linear algebraic mixing condition (about the largest singular value). Roughly speaking, linear algebraic mixing implies combinatorial mixing (as in the Expander Mixing Lemma), whereas algebraic mixing implies weaker combinatorial mixing (as in Bilu and Linial (2006)).
- Labeled examples: Both works show a lower bound of $|\mathcal{H}|^{\Omega(1)}$ on the number of labeled examples needed given bounded space.
- Space: Raz shows a stronger lower bound on the space complexity: roughly $(\log|\mathcal{H}|)^2$ rather than roughly $1.25\log|\mathcal{H}|$ in our paper. Recall that the best possible lower bound is $\log|\mathcal{H}|\log|\mathcal{X}|$, which could be much higher than $(\log|\mathcal{H}|)^2$ Kol et al. (2017).

Subsequently to those works, we were able to use the framework presented in the current work to match Raz's bound Moshkovitz and Moshkovitz (2017). We hope that our framework could be used to prove optimal results for wide families of hypotheses.

1.5. Proof Outline

We define a measure for the progress that the learner makes during the execution of the algorithm, which we call *certainty*. Certainty measures how well the learner managed to narrow down the candidates for the underlying hypothesis. The certainty is low when the algorithm starts, and should be high when the algorithm ends. Our analysis argues that when the memory is bounded the certainty cannot increase much after seeing a new labeled example. It thereby implies that the number of labeled examples that the learner sees must be large.

Assume a probability distribution over hypotheses in $\mathcal{H}$. For a memory state m at time t, let $\Pr(m)$ be the probability that the algorithm lands in m when the underlying hypothesis is drawn from the distribution and the examples are chosen at random. Let $\Pr(h|m)$ be the probability that h is the underlying hypothesis conditioned on the algorithm being in state m. We define the *certainty* of a memory m by

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

Note that a memory with low certainty is one for which many different hypotheses are possible. We define the average certainty of a set of memories M at time t by taking the weighted sum of the

individual certainties

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

When we refer to the certainty of the algorithm we typically refer to the certainty of the full set of memory states, or to the certainty of this set after the removal of a few memories (we'll explain the reason for removing memories shortly).

If the underlying hypothesis is picked uniformly from a set of $\Theta(|\mathcal{H}|)$ hypotheses, then at the start time the certainty is $O(1/|\mathcal{H}|)$. In contrast, an algorithm that identifies the underlying hypothesis with high probability must have large certainty in its final time step. We prove that assuming the memory is bounded, at each time step t – ignoring some low probability sequences of examples – there exists a high probability set of memories M_t, a large set of hypotheses H_t such that for h picked uniformly from H_t,

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

for some small constant $\epsilon > 0$. This implies that $\Omega(|\mathcal{H}|^{\epsilon})$ labeled examples are needed.

As an example, consider the *enumerator algorithm* that goes through the hypotheses in order. The algorithm maintains a current hypothesis at each time step. If the labeled example is inconsistent with the current hypothesis, the algorithm moves on to the next hypothesis. If the underlying hypothesis is one of the first few hypotheses the algorithm considers, the algorithm is likely to identify that. Moreover, the certainty of the first memory states (the ones associated with the first hypotheses) is high after sufficiently many time steps. However, if one omits the first hypotheses and memory states, the certainty is low.

In order to bound the certainty we analyze the *knowledge graph* associated with the algorithm at every time step t. The knowledge graph is a bipartite graph on memory states and hypotheses, defined as follows.

Definition 3 (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 the start time the knowledge graph always has one memory state that is connected to all hypotheses. We think of such a knowledge graph as "expanding". Formally, we define a non-standard expansion property that we name "K-expander" ("K" is for "knowledge graph") that applies to this knowledge graph.

Definition 4 *We say that a distribution p over the memories is β-enlarging if for every memory m it holds that $p(m) \leq \frac{\Pr(m)}{\beta}$.*

Definition 5 (K-expander) *A knowledge graph $G_t = (\mathcal{H}, \mathcal{M}, E_t)$ is an $(\alpha, \beta, \epsilon)$-K-expander if for any $S \subseteq \mathcal{H}$, $|S| \geq \alpha |\mathcal{H}|$, and any β-enlarging distribution T over the memories*

$$\Pr(S|T) \leq \frac{|S|}{|\mathcal{H}|} + \epsilon.$$

At the first time step, there is just one enlarging distribution over memories: the one that picks the starting memory state with probability 1. We show that the knowledge graph of any algorithm for a mixing hypothesis class $\mathcal{H}$ at any early enough time step is K-expanding with small ϵ. In contrast, the knowledge graph of an algorithm that successfully identified the underlying hypothesis is not K-expanding.

We use an inductive argument to analyze for every time t:

1. The K-expansion of the knowledge graph.

2. The certainty of the algorithm.

Towards 1 we show that a K-expanding knowledge graph of a learner with low certainty $cer^t(M_t)$ remains K-expanding after the $(t+1)$'th step. Towards 2 we show that K-expansion at time $t+1$ prevents the learner from increasing the certainty at time $t+1$. We discuss these proofs next.

Preservation of K-expansion: Fix a large set of possible hypotheses $S \subseteq \mathcal{H}$, $|S| \geq \alpha |\mathcal{H}|$, and a β-enlarging distribution T over the memories at time $t+1$. Note that T induces a distribution T' over the memories at time t that is β-enlarging as well. Moreover, the probability of the labeled examples leading from T' to T has to be large. From the K-expansion at time t, we know that $\Pr(S|T') \leq |S| / |\mathcal{H}| + \epsilon$. The challenge is to argue that the example seen after time t does not reveal much information about the underlying hypothesis and its membership in S. Concretely, we'd like to show that $\Pr(S|T)$ is not much larger than $\Pr(S|T')$. Since the certainty at time t is low, we can focus only on time-t memory states m for which there are many possible underlying hypotheses. For such memory states, because of the mixing property of the hypothesis class, there can be only a small fraction of "bad" labeled examples that reveal much information about the underlying hypothesis. Since the probability of labeled examples leading from T' to T has to be large, the probability of "bad" examples is low even within those.

Certainty remains low: Since the memory size is bounded, a typical memory state at time $t+1$ has many sources, i.e., large in-degree in the branching program. We consider two extreme cases:

Heavy source: There is a large set of possible labeled examples $S \subseteq \mathcal{X}$ such that the algorithm progresses from one memory m' at time t to a memory m at time $t+1$ if it is given any labeled example taken from S. For instance, the enumerator algorithm we discussed above has heavy sources: each time $t+1$ memory m has two memories at time t that lead to it, the one associated with the same hypothesis, m'_1 and the one associated with the previous hypothesis, m'_2. For example, m'_1 is connected to m with $|\mathcal{X}| / 2$ labeled examples.

The case of heavy sources is the simplest to handle, and does not require any assumptions about the K-expansion of the knowledge graph, only the assumption of low certainty at time t. The idea is to focus on time-t memory states with low certainty, i.e., those that have many possible hypotheses. For such memory states that transition to a time-$(t+1)$ memory state via any one of many labeled examples, the mixing property of the hypothesis class implies that most possible time-t hypotheses are still viable for the time-$(t+1)$ memory state. In other words, the time-$(t+1)$ memory state has low certainty as well.

Many source: Here there is a large number of time t memories M that lead to one memory m at time $t+1$. For instance, the memory states of an algorithm that stores only the last labeled example have many sources.

The K-expansion of the knowledge graph ensures that the time-$(t+1)$ memory state m receives no substantial information about the underlying hypothesis h from the time-t memories leading to

it. The challenge is to account for the information that is received from the example seen after time t. Roughly speaking, a full bit of information about h may be deduced from the example, and

$$\Pr(h|m) \leq 2.2 \Pr(h|M).$$

We use the low certainty at time t and the K-expansion to argue that the "confusion" due to the many different sources M compensates for the information received from the example.

Every time $t+1$ memory can have both heavy and many sources. We show how to decompose almost all of the sources to either heavy or many, combining both analyses to argue that the certainty does not increase substantially.

2. Preliminaries

2.1. Probability

Claim 1 *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 $\sum_{i\in A'} p(i) \leq \sqrt{|A'|r}$.*

Proof Using Jensen's inequality:

$$\begin{aligned}\left(\frac{1}{|A'|}\sum_{i\in A'} p(i)\right)^2 &\leq \frac{1}{|A'|}\sum_{i\in A'} p(i)^2 \\ &\leq \frac{r}{|A'|}.\end{aligned}$$

Or equivalently,

$$\sum_{i\in A'} p(i) \leq |A'|\sqrt{\frac{r}{|A'|}} = \sqrt{|A'|r}.$$

■

Claim 2 (generalized law of total probability) *For any events A, B and a partition of the sample space $C_1, \ldots, C_n$,*

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

Proof

$$\begin{aligned}\sum_i \Pr(A|B,C_i)\Pr(C_i|B) &= \sum_i \frac{\Pr(A,B,C_i)}{\Pr(B,C_i)}\frac{\Pr(C_i,B)}{\Pr(B)} \\ &= \frac{1}{\Pr(B)}\sum_i \Pr(A,B,C_i) \\ &= \frac{\Pr(A,B)}{\Pr(B)} = \Pr(A|B)\end{aligned}$$

■

Claim 3 (generalized Bayes' theorem) *For any three events A, B, C,*

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

Proof

$$\begin{aligned}\Pr(B|A,C)\frac{\Pr(A|C)}{\Pr(B|C)} &= \frac{\Pr(B,A,C)}{\Pr(A,C)}\frac{\Pr(A,C)\Pr(C)}{\Pr(C)\Pr(B,C)} \\ &= \frac{\Pr(B,A,C)}{\Pr(B,C)} \\ &= \Pr(A|B,C)\end{aligned}$$

■

Let us prove a simple claim that states that the probability of event conditioning on a set of disjoint events is actually a weighted sum.

Claim 4 *Suppose $B_1, \ldots, B_n$ are some disjoint events. Then,*

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

Proof

$$\begin{aligned}\Pr(A|B_1 \cup \ldots \cup B_n) &= \frac{\Pr(A \cap (B_1 \cup \ldots \cup B_n))}{\Pr(B_1 \cup \ldots \cup B_n)} \\ &= \frac{\Pr((A \cap B_1) \cup \ldots \cup (A \cap B_n))}{\Pr(B_1 \cup \ldots \cup B_n)} \\ &= \frac{\sum_{i=1}^{n} \Pr(A \cap B_i)}{\Pr(B_1 \cup \ldots \cup B_n)} \\ &= \sum_{i=1}^{n} \Pr(A|B_i)\frac{\Pr(B_i)}{\Pr(B_1 \cup \ldots \cup B_n)}\end{aligned}$$

■

2.2. 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*

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

Claim 5 (Union of mixing graphs is mixing) *If (A, B_1, E_1) is* d*-mixing with average left degree d_1 and (A, B_2, E_2) is* d*-mixing with average left degree d_2 and $|B_1| = |B_2|$ then $(A, B_1 \cup B_2, E_1 \cup E_2)$ is $2\mathrm{d} + \sqrt{\frac{|A|}{|B_1 \cup B_2|}}|d_1 - d_2|$-mixing.*

Proof Fix $S \subseteq A, T \subseteq B_1 \cup B_2$, and denote $T_1 = B_1 \cap T$ and $T_2 = B_2 \cap T$. Notice that the average left degree in the new graph is $d_1 + d_2$ and $|B_1 \cup B_2| = 2|B_1| = 2|B_2|$.

$$\begin{aligned}
\left| |E(S,T)| - \frac{|S||T|}{|B_1 \cup B_2|/d_1+d_2} \right| &= \left| |E(S,T_1)| + |E(S,T_2)| - \frac{|S||T_1|}{2|B_1|/d_1+d_2} - \frac{|S||T_2|}{2|B_2|/d_1+d_2} \right| \\
&\leq \left| |E(S,T_1)| - \frac{|S||T_1|}{|B_1|/d_1} \right| + \left| |E(S,T_1)| - \frac{|S||T_2|}{|B_2|/d_2} \right| + \\
&\quad \left| \frac{|S||T_1|}{|B_1|/d_1} - \frac{|S||T_1|}{2|B_1|/d_1+d_2} \right| + \left| \frac{|S||T_2|}{|B_2|/d_2} - \frac{|S||T_2|}{2|B_2|/d_1+d_2} \right| \\
&\leq \mathrm{d}\sqrt{|S||T_1|} + \mathrm{d}\sqrt{|S||T_2|} + \\
&\quad \frac{|S||T_1|}{|B_1|} \cdot \left| d_1 - \frac{d_1+d_2}{2} \right| + \frac{|S||T_2|}{|B_2|} \left| d_2 - \frac{d_1+d_2}{2} \right| \\
&\leq 2\mathrm{d}\sqrt{|S||T|} + \frac{|S||T|}{|B_1|} \frac{|d_1 - d_2|}{2} \\
&\leq \left(2\mathrm{d} + \sqrt{\frac{|A|}{2|B_1|}} |d_1 - d_2| \right) \sqrt{|S||T|}
\end{aligned}$$

■

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

Claim 6 (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 $\epsilon > 0$.*

Proof Fix $T \subseteq B$. Define $S_1 = \{a \in A | \frac{|\Gamma(a) \cap T|}{d_A} - \frac{|T|}{|B|} > \epsilon\}, S_2 = \{a \in A | \frac{|T|}{|B|} - \frac{|\Gamma(a) \cap T|}{d_A} > \epsilon\}$. Let us bound the size of each of these sets:

$$|S_1| d_A \left(\frac{|T|}{|B|} + \epsilon \right) < |E(S_1, T)| \leq \frac{|S_1||T|}{|B|/d_A} + \mathrm{d}\sqrt{|S_1||T|},$$

where the right inequality follows from the mixing property and the left inequality follows from the definition of S_1. This means that

$$|S_1| d_A \epsilon < \mathrm{d}\sqrt{|S_1||T|}$$

$$|S_1| < \frac{\mathrm{d}^2 |T|}{d_A^2 \epsilon^2} \Rightarrow \frac{|S_1|}{|A|} \leq \frac{\mathrm{d}^2 |B|}{d_A^2 \epsilon^2 |A|}$$

Similarly for S_2,

$$\frac{|S_2||T|}{|B|/d_A} - \mathrm{d}\sqrt{|S_2||T|} \leq |E(S_2, T)| < |S_2| d_A \left(\frac{|T|}{|B|} - \epsilon \right) \Rightarrow \frac{|S_2|}{|A|} < \frac{\mathrm{d}^2 |B|}{d_A^2 \epsilon^2 |A|}$$

■

We will use the previous claim with $d_A = |B|/2$ and thus we will know that the bipartite graph is an $(\epsilon, \frac{8\mathrm{d}^2}{|B||A|\epsilon^2})$-sampler for any $\epsilon > 0$.

3. Hypotheses Graph

The hypotheses graph associated with a 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 $\{0,1\}^n$, except the zero vector and the labeled examples are $\{0,1\} \times \{0,1\}^n$ (i.e., $|\mathcal{H}| = 2^n$ and $|\mathcal{X}| = 2 \cdot 2^n$).

Lemma 8 (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 5 imply that the hypotheses graph of $PARITY(n)$ is $\sqrt{|\mathcal{H}|}$-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 Krivelevich and Sudakov (2006)).

We can rephrase Claim 6 for the hypotheses graph and get

Proposition 9 *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 $\epsilon > 0$.*

3.1. 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 10 (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 11 *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 $\alpha, \beta \in (0, 1)$.*

In the next claim we will prove that the the degree of each vertex is similar ($\sim \frac{|A|d_A}{|B|}$).

Claim 7 (near regularity) *Suppose that a bipartite graph (A, B, E) is an $(\alpha, \beta, \epsilon)$-H-expander, then except for $2\beta|B|$ vertices in B, the degree of the vertex is in $[|A|(d_A/|B| - \epsilon), |A|(d_A/|B| + \epsilon)]$.*

Proof Define the two sets

$$T_1 = \{b \in B | |E(A,b)| < |A|(d_A/|B| - \epsilon)\}, \quad T_2 = \{b \in B | |E(A,b)| > |A|(d_A/|B| + \epsilon)\}.$$

We will prove that the size of each of these sets is at most $\beta|B|$. By the definition of T_1 we know that

$$|E(A, T_1)| < |A|(d_A/|B| - \epsilon)|T_1|.$$

By the H-expander property, if $|T_1| \geq \beta|B|$ then

$$|A||T_1| \left(\frac{d_A}{|B|} - \epsilon \right) \leq |E(A, T_1)|,$$

and we get a contradiction. Similar argument also holds for T_2. ■

4. The Correct Hypothesis Must be Returned

A PAC learner needs only to find an approximation of the underlying hypothesis h^*. In other words, if $\mathcal{D}$ is the underlying distribution over the labeled examples $\mathcal{X}$, then the learning algorithm should return an hypothesis h that agree with h^* with high probability over $\mathcal{D}$. A PAC learner should return such an hypothesis for any $\mathcal{D}$, specifically when $\mathcal{D}$ is the uniform distribution. In this case h should agree with h^* on most of the examples. In this section we prove that a PAC learner for a mixing hypothesis class must in fact identify the underlying hypothesis *exactly* (with high probability).

To show this, we point to a large number of hypotheses that are all far from one another. We do so in two steps. First, we show in the next claim that for each hypothesis, number of hypotheses that agree with it on at least $3/4$ of the examples is small. Then we use Turán's theorem to prove that in such a case there must be a large subset of hypotheses that are all far from each other.

Claim 8 *If a bipartite (A, B, E) with average left degree $\bar{d}_A$ is d-mixing then for every set $T \subseteq B$, the number of vertices $a \in A$ with $|E(a,T)| \geq \frac{1.5|T|}{|B|/d_A}$ is at most*

$$\frac{\mathrm{d}^2}{|T|} \cdot \left(\frac{2|B|}{\bar{d}_A}\right)^2$$

Proof Denote $S = \{a \in A || E(a,T)| \geq \frac{1.5|T|}{|B|/d_A}\}$. This implies that $|E(S,T)| \geq \frac{1.5|S||T|}{|B|/\bar{d}_A}$. From the definition of d-mixing we know that

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

Combining these two inequalities,

$$\frac{1.5|S||T|}{|B|/\bar{d}_A} \leq \frac{|S||T|}{|B|/\bar{d}_A} + \mathrm{d}\sqrt{|S||T|}$$

This means

$$\frac{|S||T|}{2|B|/\bar{d}_A} \leq \mathrm{d}\sqrt{|S||T|}$$

Or, in other words,

$$|S| \leq \frac{\mathrm{d}^2}{|T|} \cdot \left(\frac{2|B|}{\bar{d}_A}\right)^2$$

■

Lemma 12 (Turán's theorem) *Let G be any graph with n vertices without a $r+1$-clique, then the number of edges in G is at most*

$$\left(1 - \frac{1}{r}\right) \cdot \frac{n^2}{2}$$

See Aigner (1995) for more details.

Claim 9 *If the hypotheses graph $(\mathcal{H}, \mathcal{X}, E)$ is d-mixing, then there is a subset of hypotheses $H \subseteq \mathcal{H}$, with $|H| \geq \frac{|\mathcal{H}|}{1+\frac{16\mathrm{d}^2}{|\mathcal{X}|}}$, such that every two hypotheses $h_1, h_2 \in H$ have agreement less than $3/4$, i.e., $|\{x|h_1(x) = h_2(x)\}| \leq \frac{3}{4} \cdot \frac{|\mathcal{X}|}{2}$.*

Proof For any hypothesis h, use the previous claim with T equal to all the neighbors of h (there are $|\mathcal{X}|/2$ such neighbors). Thus, only $\frac{32\mathrm{d}^2}{|\mathcal{X}|}$ of the hypotheses agree with h on at least $\frac{3}{4}$ of the examples. For the sake of the proof, create a graph with vertices $\mathcal{H}$ that are connected if they agree on less than $3/4$ fraction of the examples. To prove the claim, we need to find a large clique in the new graph.

The number of edges in the graph is at least $\binom{|\mathcal{H}|}{2} - 32\mathrm{d}^2$. Using Turán's theorem, if there is no clique of size $r+1$, then

$$\left(1 - \frac{1}{r}\right) \cdot \frac{|\mathcal{H}|^2}{2} \geq \binom{|\mathcal{H}|}{2} - 32\mathrm{d}^2 \frac{|\mathcal{H}|}{|\mathcal{X}|}.$$

$$\left(1-\frac{1}{r}\right) \geq \frac{|\mathcal{H}|(|\mathcal{H}|-1)-16\mathrm{d}^2\frac{|\mathcal{H}|}{|\mathcal{X}|}}{|\mathcal{H}|^2}$$

$$\left(1-\frac{1}{r}\right) \geq \frac{|\mathcal{H}|-1-\frac{16\mathrm{d}^2}{|\mathcal{X}|}}{|\mathcal{H}|}$$

$$\frac{1+\frac{16\mathrm{d}^2}{|\mathcal{X}|}}{|\mathcal{H}|} \geq \frac{1}{r}$$

$$r \geq \frac{|\mathcal{H}|}{1+\frac{16\mathrm{d}^2}{|\mathcal{X}|}}$$

■

5. 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 of 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 a 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 state $m : q_t(m) = \sum_h q_t(m|h)q_t(h)$. 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, $q_t(h|m) = \frac{q_t(m|h)q_t(h)}{q_t(m)}$.

5.1. K-expander

In a later section we will prove that the knowledge graph preserves a pseudo-random property. To this proof a stronger notion than sampler is needed. Specifically, not only sets should sample well a set of hypotheses but also a large set of distributions. Notice that the knowledge graph can be highly irregular, e.g., there can be few memories connected to most hypotheses, whereas other memories may not be connected to any hypothesis. Our definitions are tailored to irregular graphs.

Definition 14 (β-enlarging) *Let $0 < \beta \leq 1$. We say that a distribution p over memories is β-enlarging with respect to a probability distribution q over memories if for every memory m, $p(m) \leq \frac{q(m)}{\beta}$.*

Any distribution is β-enlarging for sufficiently small β. Only q is $1-$enlarging. When q is uniform the definition merges with the definition of min-entropy.

For any β-enlarging distribution p with respect to a distribution q,

$$1 = \sum_{m|p(m)\neq 0} p(m) \leq \sum_{m|p(m)\neq 0} q(m)/\beta \Rightarrow \sum_{m|p(m)\neq 0} q(m) \geq \beta,$$

this means that the support of a β-enlarging distribution defines a set of memories with q-weight at least β.

Definition 15 (K-expander) *The knowledge graph $(\mathcal{H}, \mathcal{X}, E)$ is an $(\alpha, \beta, \epsilon) -$ K-expander if for any $S \subseteq \mathcal{H}$ with $|S| \geq \alpha|\mathcal{H}|$ and a β-enlarging distribution T it holds that*

$$\Pr(S|T) \leq \frac{|S|}{|\mathcal{H}|} + \epsilon.$$

(K in K-expander stands for Knowledge graph).

In the rest of the section we prove that the knowledge graph at time $t = 1$ is a K-expander. We can assume without loss of generality that the knowledge graph after the first example is the hypotheses graph (since the algorithm can save in memory the first example). From Proposition 11 we know that the hypotheses graph is a $(\alpha_1, \beta_1, \epsilon_1) -$ H-expander, for any $\alpha_1, \beta_1 \in (0, 1)$ and $\epsilon_1 = \frac{2\mathrm{d}}{\sqrt{\alpha_1|\mathcal{H}|\beta_1|\mathcal{X}|}}$. Later (in Section 11) we will choose α_1, β_1.

Definition 16 *Define M_1 to be all memories m (i.e., examples) with degree*

$$|\mathcal{H}|(1/2 - \epsilon_1) \leq d_m \leq |\mathcal{H}|(1/2 + \epsilon_1).$$

We remark that using Claim 7, M_1 must be large.

Claim 10 *If the hypotheses graph is d-mixing and $\epsilon_1 \leq 1/4$, then G_1' is an*

$$(\alpha_1, \beta_1, 8\epsilon_1 + \alpha_1) - \textit{K-expander}.$$

Proof

To show that the hypotheses graph is a K-expander, fix $H \subseteq \mathcal{H}$ and a β-enlarging distribution p. Denote

$$Err(H) = \{x| \Pr(H|x) > \frac{|H|}{|\mathcal{H}|} + \epsilon\}$$

(we pick ϵ later). From the definition of M_1 and $Err(H)$ we know that

$$|E(H, Err(H))| > \left(\frac{|H|}{|\mathcal{H}|} + \epsilon\right)|\mathcal{H}|(1/2 - \epsilon_1)|Err(H)|$$

The right term is equal to

$$\frac{|H||Err(H)|}{2} - |H||Err(H)|\epsilon_1 + \epsilon|\mathcal{H}|(1/2 - \epsilon_1)|Err(H)|$$

From the mixing property (Definition 6) we know that

$$\frac{|Err(H)||H|}{2} + \mathrm{d}\sqrt{|Err(H)||H|} \geq |E(H, Err(H))|.$$

Combining the last two inequalities we get

$$\frac{|Err(H)||H|}{2} + \mathrm{d}\sqrt{|Err(H)||H|} > \frac{|H||Err(H)|}{2} - |H||Err(H)|\epsilon_1 + \epsilon|\mathcal{H}|(1/2-\epsilon_1)|Err(H)|$$

$$\mathrm{d}\sqrt{|Err(H)||H|} > (\epsilon|\mathcal{H}|(1/2-\epsilon_1) - |H|\epsilon_1)|Err(H)|$$

$$\frac{\mathrm{d}\sqrt{|H|}}{(\epsilon|\mathcal{H}|(1/2-\epsilon_1) - |H|\epsilon_1)} > \sqrt{|Err(H)|}$$

$$\frac{\mathrm{d}^2|H|}{(\epsilon|\mathcal{H}|(1/2-\epsilon_1) - |H|\epsilon_1)^2} > |Err(H)|$$

The maximal value of the left term is, using $\epsilon_1 \leq 1/4$,

$$\frac{\mathrm{d}^2}{|\mathcal{H}|(\epsilon/4-\epsilon_1)^2}$$

We need to bound

$$\begin{aligned}
\sum_{x\in\mathcal{X}} \Pr(H|x)p(x) &= \sum_{x\notin Err(H)} \Pr(H|x)p(x) + \sum_{x\in Err(H)} \Pr(H|x)p(x) \\
&\leq \sum_{x\notin Err(H)} \left(\frac{|H|}{|\mathcal{H}|} + \epsilon\right) p(x) + \sum_{x\in Err(H)} 1\cdot p(x) \\
&\leq \frac{|H|}{|\mathcal{H}|} + \epsilon + \sum_{x\in Err(H)} \frac{q_1(x)}{\beta} \\
\text{(definition of } q_1) \quad &\leq \frac{|H|}{|\mathcal{H}|} + \epsilon + \sum_{x\in Err(H)} \frac{|\mathcal{H}|(1/2+\epsilon_1)}{|\mathcal{H}|(1/2-\epsilon_1)|\mathcal{X}|\beta} \\
&\leq \frac{|H|}{|\mathcal{H}|} + \epsilon + \frac{\mathrm{d}^2}{|\mathcal{H}|(\epsilon/4-\epsilon_1)^2} \cdot \frac{(1/2+\epsilon_1)}{(1/2-\epsilon_1)|\mathcal{X}|\beta} \\
(\epsilon_1 \leq 1/4) \quad &\leq \frac{|H|}{|\mathcal{H}|} + \epsilon + \frac{\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|(\epsilon/4-\epsilon_1)^2} \cdot \frac{(1+8\epsilon_1)}{\beta}
\end{aligned}$$

Take $\epsilon = 8\epsilon_1$ and notice that by the definition of ϵ_1,

$$\frac{\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|\epsilon_1^2} = \frac{\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|\left(\frac{2\mathrm{d}}{\sqrt{\alpha_1|\mathcal{H}|\beta_1|\mathcal{X}|}}\right)^2} = \frac{\alpha_1\beta_1}{4}$$

Thus, the term we would like to bound is at most $8\epsilon_1 + \frac{\alpha_1\beta_1}{4}\cdot\frac{3}{\beta_1} < 8\epsilon_1 + \alpha_1$. ■

6. 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, $H_1 = \mathcal{H}$ and M_1 is as defined in Definition 16. 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 17 (certainty) *The* certainty *of a memory m at time t is defined as*

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

The average certainty *of the 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 its 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 its 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,

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

for some $c > 0$, that is of the order $|\mathcal{H}|^\epsilon$, for some small constant ϵ. Oftentimes, we will omit c when it is clear from the context. For all $t \geq 1$ we will make sure that M_t will not include Bad_M^c (and additional memories, as will be defined in later sections). The next claim proves that removing bad memories does not reduce too much the weight.

Claim 11 *For any $c > 0$ and time t, $q_t(Bad_M^c) \leq 1/c$*

Proof We can define a probability distribution over M using q_t in the following way. For each $m \in M$ its probability is defined by $\frac{q_t(m)}{q_t(M)}$. The assumption in the claim states that

$$\begin{aligned}\mathbb{E}_{m^t \in M}[\sum_h q_t(h|m^t)^2] &= \sum_{m^t \in M} \frac{q_t(m^t)}{q_t(M)} \sum_h q_t(h|m^t)^2 \\ &= cer^t(M)/q_t(M).\end{aligned}$$

Using Markov's inequality we know that the probability of

$$Bad_M^c = \left\{ m^t \in M \Big| \sum_h q_t(h|m^t)^2 > (c \cdot q_t(M)) \cdot \frac{cer^t(M)}{q_t(M)} \right\},$$

is at most $1/(c \cdot q_t(M))$, i.e.,

$$\sum_{m^t \in Bad_M} \frac{q_t(m^t)}{q_t(M)} \leq \frac{1}{q_t(M)c} \Rightarrow \frac{q_t(Bad_M)}{q_t(M)} \leq \frac{1}{q_t(M)c} \Rightarrow q_t(Bad_M) \leq 1/c.$$

■

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

Claim 12 *For each memory m, hypothesis h and time t*

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

Proof

$$\begin{aligned}q_t(m)q_t(h|m)^2 &= q_t(m)q_t(h|m)q_t(h|m) \\ \text{(by Bayes' theorem)} &= q_t(m)q_t(h|m)\frac{q_t(m|h)q_t(h)}{q_t(m)} \\ &= q_t(h)q_t(h|m)q_t(m|h)\end{aligned}$$

■

In particular we can prove

Claim 13 *The average certainty is also equal to*

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

Proof

$$\begin{aligned} cer^t(M) &= \sum_{m\in M} q_t(m)\left(\sum_{h\in\mathcal{H}} q_t(h|m)^2\right) \\ &= \sum_{m\in M,h\in\mathcal{H}} q_t(m)q_t(h|m)q_t(h|m) \\ \text{(by Bayes' theorem)} &= \sum_{m\in M,h\in\mathcal{H}} q_t(m)q_t(h|m)\frac{q_t(m|h)q_t(h)}{q_t(m)} \\ &= \sum_{m\in M,h\in\mathcal{H}} q_t(h|m)q_t(m|h)q_t(h) \\ &= \sum_{h\in\mathcal{H}} q_t(h)\sum_{m\in M} q_t(h|m)q_t(m|h) \end{aligned}$$

■

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

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

Given the last claim in mind we define

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

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

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

Claim 14 *For any $c > 0$, $|H_{t+1}| \geq (1 - 1/c)|H_t|$.*

Proof From Claim 13 and from Markov's inequality, we know that

$$\Pr_{h\sim q_t}(h \in Bad_H) \leq 1/c$$

Since $\Pr_{h\sim q_t}(h \in Bad_H) = \frac{|Bad_H|}{|H_t|}$ we get that

$$|Bad_H| \leq |H_t|/c.$$

Thus,

$$|H_{t+1}| \geq |H_t| - |Bad_H| \geq (1 - 1/c)|H_t|.$$

■

In the rest of the paper we will prove that the average certainty of M_t, even for a large $t \sim |\mathcal{H}|^{\Omega(1)}$, will be at most $\frac{3}{|\mathcal{H}|}$.

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 15 *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*

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

Proof By definition, for any $m \in M_t$ it holds that $\sum_h q_t(h|m)^2 \leq c \cdot cer^t(M_t)$. This implies that for any h, $q_t(h|m) \leq \sqrt{c \cdot cer^t(M_t)}$.

Each memory m is associated with some hypothesis $h(m)$ that the algorithm returns as its answer when reaching m. The probability that the algorithm returns the correct hypothesis, assuming the true hypothesis is h, is at most

$$\sum_{m|h(m)=h} q_t(m|h) + \gamma.$$

Let us explore the first term

$$\sum_{m|h(m)=h} q_t(m|h) \quad = \quad \sum_{m \in M_t|h(m)=h} q_t(m|h) + \sum_{m \notin M_t|h(m)=h} q_t(m|h).$$

Let us focus on the first term, by Bayes' theorem, it is equal to

$$\sum_{m \in M_t|h(m)=h} \frac{q_t(h|m)q_t(m)}{q_t(h)} \leq \sqrt{c \cdot cer^t(M_t)} \sum_{m \in M_t|h(m)=h} \frac{q_t(m)}{q_t(h)}.$$

We will use Markov's inequality to bound this term. Let us first calculate the following expectation

$$\sum_{h \in H_t} \left[\sum_{m \in M_t|h(m)=h} q_t(m) \right] \leq 1 \Rightarrow \mathbb{E}_{h \sim q_t} \left[\sum_{m \in M_t|h(m)=h} q_t(m) \right] \leq \frac{1}{|H_t|}$$

Thus, for at most $1/3$ fraction of the hypotheses h,

$$\sum_{m \in M_t|h(m)=h} q_t(m) \geq \frac{3}{|H_t|}.$$

In other words, for at least $2/3$ fraction of the hypotheses the first term is bounded by $3\sqrt{c \cdot cer^t(M_t)}$. As for the second term, it is bounded by

$$\sum_{m \notin M_t} q_t(m|h).$$

Averaging over all $h \in H_t$ we get

$$\frac{1}{|H_t|} \sum_h \sum_{m \notin M_t} q_t(m|h) = \sum_{m \notin M_t} \sum_h q_t(m|h)q_t(h) = \sum_{m \notin M_t} q_t(m) = 1 - q_t(M_t).$$

Thus, by Markov inequality, for at most $1/3$ fraction of the hypotheses $\sum_{m \notin M_t} q_t(m|h) \geq 3(1 - q_t(M_t))$. To sum up, there is an hypothesis for which the sum of the first and the second term is at

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

We also define a weighted certainty using a weight vector w of length $|\mathcal{M}|$ and each coordinate in w is some value in $[0, 1]$ by

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

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

7. Representative Labeled Examples

In this section we define the set of non-representative labeled examples. We then prove that this set is small and thus can be removed.

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 18 *Let m be a memory state at time t. We say that a labeled example x is* representative *at m if*

$$\frac{1}{1.1|\mathcal{X}|} \leq q_{t+1}(x|m) \leq \frac{1.1}{|\mathcal{X}|}$$

We denote the set of labeled examples that are not representative at m by $NRep(m)$, i.e.,

$$NRep(m) := \left\{ x \in \mathcal{X} | q_{t+1}(x|m) < \frac{1}{1.1|\mathcal{X}|} \right\} \cup \left\{ x \in \mathcal{X} | q_{t+1}(x|m) > \frac{1.1}{|\mathcal{X}|} \right\}.$$

The next claim will imply an equivalent definition for this set.

Claim 16 *For any set of labeled examples $S \subseteq \mathcal{X}$ and a memory m it holds that*

$$q_t(S|m) = \sum_h \Pr(S|h) q_t(h|m).$$

Proof Using Claim 2 we know that

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

■

Using Claim 16, we know that the not-representative set is also equal to

$$NRep(m) = \left\{ x \in \mathcal{X} | \sum_{h \in \mathcal{H}} \Pr(x|h) q_t(h|m) < \frac{1}{1.1|\mathcal{X}|} \right\} \cup \left\{ x \in \mathcal{X} | \sum_{h \in \mathcal{H}} \Pr(x|h) q_t(h|m) > \frac{1.1}{|\mathcal{X}|} \right\}.$$

We would like to prove that $NRep(m)$ 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 3 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 19 *We say that a left regular bipartite graph (A, B, E) is (β, ϵ) − weighted-expander with weights $w_1, \ldots, w_{|A|}$, $\sum_i w_i = 1$, $\forall i, w_i \geq 0$, and left degree d_A if for every $S \subseteq A$ and $T \subseteq B, |T| \geq \beta|B|$ it holds that*

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

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

Claim 17 *If the hypotheses graph $(\mathcal{H}, \mathcal{X}, E)$ is an $(\alpha, \beta, \epsilon)$ − H-expander and $\sum_{i=1}^{|\mathcal{H}|} w_i^2 \leq r$ then the hypotheses graph is a $(\beta, 2\epsilon + 2\sqrt{\alpha|\mathcal{H}|r})$ − weighted-expander with weights $w_1, \ldots, w_{|\mathcal{H}|}$.*

Proof Fix $S \subseteq \mathcal{H}, T \subseteq \mathcal{X}, |T| \geq \beta|\mathcal{X}|$. Denote by $Bad(T) \subseteq \mathcal{H}$ all the hypotheses that do not sample T correctly, i.e., $Bad(T) = \{h \in \mathcal{H} | \sum_{t\in T} I_{(t,h)\in E} > |T| \left(\frac{1}{2} + \epsilon\right)\}$. Then $|Bad(T)| < \alpha|\mathcal{H}|$ (because $|E(T, Bad(T))| > \frac{|Bad(T)||T|}{2} + \epsilon|Bad(T)||T|$ and if $|Bad(T)| \geq \alpha|\mathcal{H}|$ then this is a contradiction to the H-expander property).

Let us start with upper bounding $w(S,T)$.

$$\begin{aligned}
w(S,T) &= \sum_{s\in S\setminus Bad(T)} w(s) \sum_{t\in T} I_{(s,t)\in E} + \sum_{s\in S\cap Bad(T)} w(s) \sum_{t\in T} I_{(s,t)\in E} \\
&\leq \sum_{s\in S\setminus Bad(T)} w(s) \left(\frac{|T|}{2} + \epsilon|T|\right) + \sum_{s\in S\cap Bad(T)} w(s)|T| \\
&\leq \left(\frac{|T|}{2} + \epsilon|T|\right) \sum_{s\in S} w(s) + |T| \sum_{s\in Bad(T)} w(s) \\
\text{(Claim 1)} \quad &< \frac{w(S)|T|}{2} + (\epsilon + \sqrt{\alpha|\mathcal{H}|r})|T|,
\end{aligned}$$

We can lower bound $w(S,T)$ similarly. Define the set $Bad_2(T) = \{h \in \mathcal{H} | \sum_{t\in T} I_{(t,h)\in E} < |T|\left(\frac{1}{2}-\epsilon\right)\}$ and deduce that $|Bad_2(T)| < \alpha|\mathcal{H}|$ and that

$$\begin{aligned}
w(S,T) &= \sum_{s\in S\setminus Bad_2(T)} w(s)\sum_{t\in T} I_{(s,t)\in E} + \sum_{s\in S\cap Bad_2(T)} w(s)\sum_{t\in T} I_{(s,t)\in E} \\
&\geq \sum_{s\in S\setminus Bad_2(T)} w(s)\left(\frac{|T|}{2}-\epsilon|T|\right)+0 \\
\text{(Claim 1)} \quad &> \left(w(S)-\sqrt{\alpha|\mathcal{H}|r}\right)\left(\frac{|T|}{2}-\epsilon|T|\right) \\
&= w(S)\frac{|T|}{2} - \sqrt{\alpha|\mathcal{H}|r}(1/2-\epsilon)|T| - \epsilon w(S)|T| \\
&\geq w(S)\frac{|T|}{2} - \sqrt{\alpha|\mathcal{H}|r}|T| - \epsilon|T| \\
&= w(S)\frac{|T|}{2} - (\sqrt{\alpha|\mathcal{H}|r}+\epsilon)|T|
\end{aligned}$$

■

Next we will prove our main claim in this section.

Claim 18 *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, and $\sqrt{\alpha|\mathcal{H}|r}+\epsilon < 1/44$ then $|NRep(m)| \leq 2\beta$.*

Proof 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 17 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 17. So again we conclude that $|T_2| < \beta|\mathcal{X}|$. ■

8. Knowledge Graph Remains K-expander

Let us prove that a K-expander remains a K-expander even in the face of a new example, provided that the certainty is low and the hypotheses graph is mixing.

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 19 *If the hypotheses graph is an $(\alpha,\beta,\epsilon)-$H-expander and the graph G'_t is an $(\alpha',\beta',\epsilon')-$K-expander, then the graph G'_{t+1} is an $(\alpha',\beta',16\epsilon+16\sqrt{\alpha|\mathcal{H}|c\cdot cer^t(M_t)}+\frac{2\beta}{\beta'}+\epsilon')-$K-expander.*

Proof Define $\epsilon^* := 2\epsilon+2\sqrt{\alpha|\mathcal{H}|c\cdot cer^t(M_t)}$. Notice that we can assume without loss of generality that $\epsilon^* \leq 1/4$ (i.e., $\epsilon + \sqrt{\alpha|\mathcal{H}|c\cdot cer^t(M_t)} \leq 1/8$), otherwise the statement in the claim is trivial.

The distribution q_{t+1} over memories m at time $t+1$ also defines a distribution over (m^t, S) where m^t is a memory at time t and $S \subseteq \mathcal{X}$ is a set labeled examples, in the following way

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

For ease of notation, if $S = \{x\}$ (i.e., S includes only one labeled example), we simply write $q_{t+1}(m^t, x)$.

Fix a β'-enlarging distribution p (with respect to q_{t+1}) over memories at time $t+1$ and denote its support by M. For each $m \in M$, denote $p(m) = \frac{q_{t+1}(m)}{\beta'_m}$, for some $\beta'_m \geq \beta'$. This induces a distribution over (m^t, S), where m^t is a memory at time t and S is a set of labeled examples $p(m^t, S) = \frac{q_t(m^t)\Pr(S|m^t)}{\beta'_m}$.

Fix a set of hypotheses $H \subseteq \mathcal{H}$ with $|H| \geq \alpha|\mathcal{H}|$.

We will start by proving that for any m^t, memory at time t, and for any $S \subseteq \mathcal{X}, |S| \geq \beta|\mathcal{X}|$, the probability $q_{t+1}(H|m^t, S)$ is not much more than $q_t(H|m^t)$. Fix m^t, a memory at time t.

$$\begin{aligned}
q_{t+1}(H|m^t,S) &= \sum_{h\in H} q_{t+1}(h|m^t,S) \\
\text{(Claim 3)} &= \sum_{h\in H} q_t(h|m^t)\frac{q_{t+1}(S|h,m^t)}{q_{t+1}(S|m^t)} \\
\text{(Claim 2)} &= \sum_{h\in H} q_t(h|m^t)\frac{q_{t+1}(S|h,m^t)}{\sum_{h'} q_{t+1}(S|h',m^t)q_t(h'|m^t)} \\
&= \sum_{h\in H} q_t(h|m^t)\frac{\Pr(S|h)}{\sum_{h'}\Pr(S|h')q_t(h'|m^t)} \\
&= \sum_{h\in H} q_t(h|m^t)\frac{\sum_{x\in S}\Pr(x|h)}{\sum_{x\in S,h'}\Pr(x|h')q_t(h'|m^t)} \\
\text{(see below)} &= \sum_{h\in H}\frac{q_t(h|m^t)\sum_{x\in S}\frac{2}{|\mathcal{X}|}I_{(x,h)\in E}}{\sum_{h'\in\mathcal{H}} q_t(h'|m^t)\sum_{x\in S}\frac{2}{|\mathcal{X}|}I_{(x,h')\in E}}.
\end{aligned}$$

The last equality is true since if $(x,h)\in E$ then $\Pr(x|h)=\frac{2}{|\mathcal{X}|}$, else $\Pr(x|h)=0$.

To further simplify this expression we define the weights $w(h)=q_t(h|m^t)$ for each hypothesis $h\in\mathcal{H}$. Using the weight notation from Section 7 we have that

$$q_{t+1}(H|m^t,S)=\frac{\sum_{h\in H,x\in S} q_t(h|m^t)I_{(x,h)\in E}}{\sum_{h'\in\mathcal{H},x\in S} q_t(h'|m^t)I_{(x,h')\in E}}=\frac{w(H,S)}{w(\mathcal{H},S)}$$

Since $\sum_{h\in\mathcal{H}} w(h)^2 \le c\cdot cer^t(M_t)$, we know from Claim 17 with ϵ^* (recall that $\epsilon^*=2\epsilon+2\sqrt{\alpha|\mathcal{H}|c\cdot cer^t(M_t)}$ that

$$\begin{aligned}
q_{t+1}(H|m^t,S) &\le \frac{\frac{q_t(H|m^t)}{2}|S|+\epsilon^*|S|}{\frac{q_t(\mathcal{H}|m^t)}{2}|S|-\epsilon^*|S|} \\
\text{(using } q_t(\mathcal{H}|m^t)=1) &\le \frac{\frac{q_t(H|m^t)}{2}|S|+\epsilon^*|S|}{\frac{1}{2}|S|-\epsilon^*|S|} \\
\text{(divide and multiply by } |S|/2) &= \frac{q_t(H|m^t)+2\epsilon^*}{1-2\epsilon^*} \\
&= \frac{q_t(H|m^t)}{1-2\epsilon^*}+\frac{2\epsilon^*}{1-2\epsilon^*} \\
\text{(for } \epsilon^*\le 1/4) &\le q_t(H|m^t)(1+4\epsilon^*)+4\epsilon^* \\
&\le q_t(H|m^t)+8\epsilon^*.
\end{aligned}$$

For each memory m^t at time t, denote by

$$Err(m^t)=\{x\in\mathcal{X}|q_{t+1}(H|m^t,x)>q_t(H|m^t)+8\epsilon^*\}.$$

Using Claim 4, we have that

$$\begin{aligned} q_{t+1}(H|m^t, Err(m^t)) &= \sum_{x \in Err(m^t)} q_{t+1}(H|m^t, x) \frac{q_{t+1}(m^t, x)}{q_{t+1}(m^t, Err(m^t))} \\ &> (q_t(H|m^t) + 8\epsilon^*) \sum_{x \in Err(m^t)} \frac{q_{t+1}(m^t, x)}{q_{t+1}(m^t, Err(m^t))} \\ &= q_t(H|m^t) + 8\epsilon^* \end{aligned}$$

Then for all $m^t \in M_t$, since for each $S \geq \beta|\mathcal{X}|$ we know that $q_{t+1}(H|m^t, S) \leq q_t(H|m^t) + 8\epsilon^*$ and from what we have just proved, $|Err(m^t)| < \beta|\mathcal{X}|$. We will show that this implies a bound on $q_{t+1}(Err(m^t)|m^t)$.

Using Claim 2, we know that for any labeled example x,

$$\begin{aligned} q_{t+1}(x|m) &= \sum_h q_{t+1}(x|h, m) q_t(h|m) \\ &= \sum_h \Pr(x|h) q_t(h|m) \\ &\leq \sum_h \frac{2}{|\mathcal{X}|} q_t(h|m) = \frac{2}{|\mathcal{X}|}, \end{aligned}$$

Hence,

$$q_{t+1}(Err(m^t)|m^t) = \sum_{x \in Err(m)} q_{t+1}(x|m) < \beta|\mathcal{X}| \frac{2}{|\mathcal{X}|} = 2\beta$$

Let us rewrite the desired expression

$$\begin{aligned} \sum_{m \in M} q_{t+1}(H|m) p(m) &= \sum_{m \in M} q_{t+1}(H| \vee_{m^t} (m^t, S^{m^t, m})) p(m) \\ \text{(Claim 4)} \quad &= \sum_{m \in M, m^t} q_{t+1}(H|m^t, S^{m^t, m}) \frac{q_{t+1}(m^t, S^{m^t, m})}{q_{t+1}(m)} \frac{q_{t+1}(m)}{\beta'_m} \\ &= \sum_{m \in M, m^t} q_{t+1}(H|m^t, S^{m^t, m}) p(m^t, S^{m^t, m}) \\ &= \sum_{m \in M, m^t} q_{t+1}(H|m^t, \vee_{x \in S^{m^t, m}} x) p(m^t, S^{m^t, m}) \\ \text{(Claim 4)} \quad &= \sum_{m \in M, m^t} q_{t+1}(H|m^t, x) \frac{q_{t+1}(m^t, x)}{q_{t+1}(m^t, S^{m^t, m})} \frac{q_{t+1}(m^t, S^{m^t, m})}{\beta'_m} \\ &= \sum_{\substack{m \in M, m^t \\ x \in S^{m^t, m}}} q_{t+1}(H|m^t, x) p(m^t, x) \\ &= \sum_{m^t, x \in \mathcal{X}} q_{t+1}(H|m^t, x) p(m^t, x) \end{aligned}$$

To bound this desired expression we divide all the pairs of (m^t, x) depending on whether $x \in Err(m^t)$ or not. Hence, it is equal to

$$\begin{aligned}
&\sum_{\substack{m^t,\\ x\notin Err(m^t)}} q_{t+1}(H|m^t,x)p(m^t,x) + \sum_{\substack{m^t,\\ x\in Err(m^t)}} q_{t+1}(H|m^t,x)p(m^t,x) \\
\leq &\sum_{\substack{m^t,\\ x\notin Err(m^t)}} (q_t(H|m^t)+8\epsilon^*)p(m^t,x) + \sum_{\substack{m^t,\\ x\in Err(m^t)}} p(m^t,x) \\
\leq &\left(\sum_{m^t,x} q_t(H|m^t)p(m^t,x)\right) + 8\epsilon^* + \sum_{m^t} \frac{q_{t+1}(m^t, Err(m^t))}{\beta'} \\
= &\left(\sum_{m^t,x} q_t(H|m^t)p(m^t,x)\right) + 8\epsilon^* + \sum_{m^t} \frac{q_{t+1}(Err(m^t)|m^t)q_t(m^t)}{\beta'} \\
\leq &\left(\sum_{m^t,x} q_t(H|m^t)p(m^t,x)\right) + 8\epsilon^* + \sum_{m^t} \frac{2\beta q_t(m^t)}{\beta'} \\
\leq &\left(\sum_{m^t,x} q_t(H|m^t)p(m^t,x)\right) + 8\epsilon^* + \frac{2\beta}{\beta'} \\
\leq &\frac{|H|}{|\mathcal{H}|} + \epsilon' + 8\epsilon^* + \frac{2\beta}{\beta'},
\end{aligned}$$

where the last inequality follows from the assumption in the claim that the knowledge graph at time t is an $(\alpha, \beta', \epsilon') -$ K-expander. ■

9. 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 20 *The set of heavy-sourced memories at time $t+1$ is defined as*

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

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 11). 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 20 and in Claim 21. Claim 20 gives an expression for $cer^{t+1}(M_{t+1}^{heavy>b})$. To understand this expression, notice that a small variant of Claim 13 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 20 *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)

$$\begin{aligned}
(\star) \qquad q_{t+1}(h|m^{t+1}) &= q_{t+1}(h|S,m^t) \\
\text{(using Claim 3)} \quad &= q_t(S|h,m^t) \frac{q_t(h|m^t)}{\Pr(S|m^t)} \\
\text{(using } \Pr(S|h,m^t) = \Pr(S|h), \quad &= \frac{\Pr(S|h) q_t(h|m^t)}{\sum_{h'} \Pr(S|m^t,h') q_t(h'|m^t)} \\
\text{and Claim 2)} \quad &= \frac{\Pr(S|h) q_t(h|m^t)}{\sum_{h'} \Pr(S|h') q_t(h'|m^t)}
\end{aligned}$$

Note that

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

Use Claim 13 and equations $(\star)$, $(\star,\star)$ to rewrite $cer_w^{t+1}(M_{t+1}^{heavy>b} \cap M)$

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

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 21 *If the hypotheses graph is an (ϵ,ϵ')-sampler, $c\geq 4$, $|H_t|\geq|\mathcal{H}|/3$, $|H_{t+1}|\geq|H_t|(1-1/c)$, $cer^t(M_t)\leq\frac{3}{|\mathcal{H}|}$, 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*

$$b\geq\max(5\epsilon c+2c\sqrt{3\epsilon' c},12a^2\epsilon' c+\epsilon),$$

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 20, $cer_w^{t+1}(M_{t+1}^{heavy>b}\cap M)$ 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 1 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 $m^t \notin Bad_{Mt}$ 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 $cer^t(M_t) \leq \frac{3}{|\mathcal{H}|}$, this implies that $\epsilon'' \leq \sqrt{3\epsilon' c}$.

Consider two cases:

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

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

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

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

We will show that

$$\sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in Err(S)}} q_t(h)q_t(h|m^t)q_t(m^t|h)w_{(m^t,S)}\Pr(S|h) \leq 6a^2\epsilon'\cdot cer^t(M_t)$$

The left hand side is at most

$$\begin{aligned}
\text{(see below)} &\leq \sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in Err(S)\cap H_t}} q_t(h)q_t(h|m^t)q_t(m^t|h)2\frac{|S|}{|\mathcal{X}|} \\
\text{(Claim 12)} &\leq \sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in Err(S)\cap H_t}} q_t(m^t)q_t(h|m^t)^2\cdot 2\frac{|S|}{|\mathcal{X}|} \\
\text{(assumption in the claim)} &\leq \sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in Err(S)\cap H_t}} q_t(m^t)(a\cdot cer^t(M_t))^2\cdot 2\frac{|S|}{|\mathcal{X}|} \\
(cer^t(M_t)\leq\frac{3}{|\mathcal{H}|}) &\leq \sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in Err(S)}} q_t(m^t)\frac{3a^2}{|\mathcal{H}|}\cdot cer^t(M_t)\cdot 2\frac{|S|}{|\mathcal{X}|} \\
(|Err(S)|\leq\epsilon'|\mathcal{H}|) &\leq \sum_{(m^t,S)\in M_{t+1}^{heavy>b}\cap M} q_t(m^t)\frac{3a^2}{|\mathcal{H}|}\cdot cer^t(M_t)\cdot 2\frac{|S|}{|\mathcal{X}|}\cdot\epsilon'|\mathcal{H}| \\
&\leq 6a^2\epsilon'\cdot cer^t(M_t)\cdot\sum_{m^t\in M_t} q_t(m^t) \\
&\leq 6a^2\epsilon'\cdot cer^t(M_t)
\end{aligned}$$

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

$$\begin{aligned}
\left(1+\frac{2}{c}\right)\Bigg[\Bigg[\sum_{\substack{(m^t,S)\in M_{t+1}^{heavy>b}\cap M \\ h\in\mathcal{H}}} & q_t(h)q_t(h|m^t)q_t(m^t|h)w_{(m^t,S)} \\
& \cdot \left(\frac{|S|}{|\mathcal{X}|}+\epsilon\right)\left(1+\frac{2\epsilon+\epsilon''}{(b-\epsilon)(1-\epsilon'')}\right)\Bigg] \\
& + 6a^2\epsilon'\cdot cer^t(M_t)\frac{1}{(b-\epsilon)(1-\epsilon'')}\Bigg]
\end{aligned}$$

Using Claim 13, (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

$$\begin{aligned}
\left(1+\frac{2}{c}\right)\Bigg[\Bigg[\sum_{(m^t,S)\in M_{t+1}^{heavy>b}\cap M} & cer^t(m^t)w_{(m^t,S)}\frac{|S|}{|\mathcal{X}|}\left(1+\frac{\epsilon}{|S|/|\mathcal{X}|}\right)\left(1+\frac{2\epsilon+\epsilon''}{(b-\epsilon)(1-\epsilon'')}\right)\Bigg] \\
& + 6a^2\epsilon'\cdot cer^t(M_t)\frac{1}{(b-\epsilon)(1-\epsilon'')}\Bigg]
\end{aligned}$$

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 $\epsilon'' \leq \sqrt{3\epsilon' c}$. We have $5\epsilon c + 2c\sqrt{3\epsilon' c} \leq b \leq 1 \Rightarrow \epsilon'' \leq \sqrt{3\epsilon' c} \leq 0.5 \Rightarrow \frac{1}{1-\epsilon''} \leq 2$. Thus, we would like to show the bound $4\epsilon c + 2\epsilon'' c \leq b - \epsilon$, so it is enough that $5\epsilon c + 2c\sqrt{3\epsilon' c} \leq b$, 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

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

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

$$\begin{aligned}&\left(1+\frac{2}{c}\right)6a^2\epsilon'\frac{1}{(b-\epsilon)(1-\epsilon'')}\\ \text{(see Item 1 below)}\ \leq\ & \left(1+\frac{2}{c}\right)\frac{1}{c}\\ \text{(see Item 2 below)}\ \leq\ & \frac{2}{c}\end{aligned}$$

1. It suffices to show that $\frac{12a^2\epsilon'}{b-\epsilon} \leq 1/c \Leftrightarrow 12a^2\epsilon' c + \epsilon \leq b$
2. $(1+2/c)1/c = 1/c + 2/c^2$ and also $2/c^2 \leq 1/c$ for $2 \leq c$.

■

10. 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 7.

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 11). 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 21 *The set of many-sourced memories at time $t+1$ is defined as*

$$M_{t+1}^{many>\beta} = \{m^{t+1} | \exists \text{ memories } m_i^t \in M_t \text{ with } \sum_i q_t(m_i^t) \geq \beta \text{ and labeled examples } x_i \notin NRep(m_i^t) \text{ that lead to } m^{t+1}\}.$$

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 19) that for any large enough $H \subseteq \mathcal{H}$ it holds that
$$q_t(H|\psi(m^{t+1})) \leq \frac{|H|}{|\mathcal{H}|} + \epsilon'$$
(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 $h \in \mathcal{H}$,
$$q_{t+1}(h|m^{t+1}) \leq 2.2 q_t(h|\psi(m^{t+1}))$$
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 $h \in \mathcal{H}$,
$$q_{t+1}(h|m^{t+1}) \leq \frac{2.3}{|\mathcal{H}|}.$$
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 22 *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*

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

Proof

$$\begin{aligned}
q_{t+1}(h|m^{t+1}) &= q_t(h|\vee_i (m_i^t, S_i)) \\
\text{(Conditional probability dfn.)} &= \frac{q_t\big(h \wedge (\vee_i (m_i^t, S_i))\big)}{q_t(\vee_i (m_i^t, S_i))} \\
\text{(De Morgan's law)} &= \frac{q_t\big(\vee_i (h \wedge (m_i^t, S_i))\big)}{q_t(\vee_i (m_i^t, S_i))} \\
\text{(Disjoint events)} &= \frac{\sum_i q_t\big(h \wedge (m_i^t, S_i)\big)}{\sum_i q_t(m_i^t, S_i)} \\
\text{(Conditional probability dfn.)} &= \frac{\sum_i q_t(h|m_i^t, S_i)\Pr(m_i^t, S_i)}{\sum_i q_t(S_i|m_i^t)\Pr(m_i^t)} \\
\text{(Claim 3 \& } q_t(S_i|h, m_i^t) = \Pr(S_i|h)) &= \frac{\sum_i \Pr(S_i|h)\frac{q_t(h|m_i^t)}{q_t(S_i|m_i^t)} q_t(S_i|m_i^t) q_t(m_i^t)}{\sum_i q_t(S_i|m_i^t)\Pr(m_i^t)} \\
&= \frac{\sum_i \Pr(S_i|h) q_t(h|m_i^t) q_t(m_i^t)}{\sum_i q_t(S_i|m_i^t) q_t(m_i^t)}
\end{aligned}$$

■

Now we are ready to prove step 2.

Claim 23 *If $m^{t+1} \in M_{t+1}^{many>\beta}$ then for any $h \in \mathcal{H}$ it holds that*

$$q_{t+1}(h|m^{t+1}) \leq 2.2 \cdot q_t(h|\psi(m^{t+1})).$$

Proof We will use the fact that if $m^{t+1} \in M_{t+1}^{many>\beta}$, 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 22 with $S_i = \{x_i\}$ for all i we know that

$$\begin{aligned}
q_{t+1}(h|m^{t+1}) &= \frac{\sum_i \Pr(x_i|h) q_t(h|m_i^t) q_t(m_i^t)}{\sum_i q_t(x_i|m_i^t) q_t(m_i^t)} \\
\text{(see below)} &\leq \frac{\sum_i \frac{2}{|\mathcal{X}|} q_t(h|m_i^t) q_t(m_i^t)}{\sum_i q_t(x_i|m_i^t) q_t(m_i^t)} \\
\text{(definition of } NRep(m_i^t)) &\leq \frac{\sum_i \frac{2}{|\mathcal{X}|} q_t(h|m_i^t) q_t(m_i^t)}{\sum_i \frac{1}{1.1|\mathcal{X}|} q_t(m_i^t)} \\
&= 2.2 \cdot \frac{\sum_i q_t(h|m_i^t) q_t(m_i^t)}{\sum_i q_t(m_i^t)} \\
&= 2.2 \cdot \sum_i q_t(h|m_i^t) \frac{q_t(m_i^t)}{q_t(\psi(m_i^{t+1}))} \\
\text{(by Claim 4)} &= 2.2 \cdot q_t(h|\psi(m^{t+1}))
\end{aligned}$$

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 24 *If the graph G'_t is an $(\alpha', \beta', \epsilon') - K$-expander, and $22\epsilon' \leq \alpha'$, then for every memory $m^{t+1} \in M_{t+1}^{many>\beta'}$ there is a set $T \subset \mathcal{H}, |T| \leq \alpha'|\mathcal{H}|$, such that for any $h \notin T$ it holds that*

$$q_{t+1}(h|m^{t+1}) \leq \frac{2.3}{|\mathcal{H}|}.$$

Proof Define $T = \{h|\frac{2.3}{|\mathcal{H}|} < q_{t+1}(h|m^{t+1})\}$, then

$$2.3\frac{|T|}{|\mathcal{H}|} < q_{t+1}(T|m^{t+1}),$$

From Claim 23 we know that for every $h \in \mathcal{H}$ it holds that

$$q_{t+1}(h|m^{t+1}) \leq 2.2 \cdot q_t(h|\psi(m^{t+1})).$$

The last two inequalities imply that

$$2.3\frac{|T|}{|\mathcal{H}|} < 2.2 \cdot q_t(T|\psi(m^{t+1})).$$

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

$$2.3\frac{|T|}{|\mathcal{H}|} < 2.2\frac{|T|}{|\mathcal{H}|} + 2.2\epsilon',$$

or in other words

$$\frac{1}{22}\frac{|T|}{|\mathcal{H}|} < \epsilon',$$

which is a contradiction since by the assumption in the claim we know that $22\epsilon' \leq \alpha'$. ■

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 25 *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

$$\begin{aligned} q_t(m)q_t(h|m)^2 &= q_t(m)q_t(h|m_1 \vee \ldots \vee m_l)^2 \\ \text{(using Claim 4)} &= q_t(m)\left(\sum_i q_t(h|m_i)\frac{q_t(m_i)}{q_t(m)}\right)^2 \\ \text{(by Jensen's inequality)} &\leq q_t(m)\sum_i \left(q_t(h|m_i)^2\frac{q_t(m_i)}{q_t(m)}\right) \\ &= \sum_i q_t(m_i)q_t(h|m_i)^2 \end{aligned}$$

■

Using Claim 12, the last claim imply the following

Corollary 22 *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 26 *If the hypotheses graph is an $(\alpha, \beta, \epsilon) - H$-expander, the graph G'_t is an $(\alpha', \beta', \epsilon') - K$-expander, $22\epsilon' \leq \alpha'$, $c \geq 45$, $cer^t(M_t) \leq \frac{3}{|\mathcal{H}|}$, $|H_{t+1}| \geq (1-1/c)|H_t|$, $3|H_t| \geq |\mathcal{H}|$, 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*

$$cer_w^{t+1}(M_{t+1}^{many>\beta'} \cap M) \leq \left[\frac{2.3}{|\mathcal{H}|} \cdot \sum_{m \in M_{t+1}^{many>\beta'} \cap M} q_{t+1}(m)w_m\right] + 13.5\alpha' a^2 cer^t(M_t)$$

Proof Using Claim 24 for every memory $m^{t+1} \in M_{t+1}^{many>\beta'}$ 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

$$q_{t+1}(h|m^{t+1}) \leq \frac{2.3}{|\mathcal{H}|}.$$

Using Claim 13,

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

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

$$\begin{aligned}
&\sum_{\substack{m^{t+1} \in M_{t+1}^{many > \beta'} \cap M \\ h \notin T_{m^{t+1}}}} q_{t+1}(h) \cdot \frac{2.3}{|\mathcal{H}|} \cdot q_{t+1}(m^{t+1}|h) w_{m^{t+1}} \\
\leq \; & \frac{2.3}{|\mathcal{H}|} \cdot \sum_{m^{t+1} \in M_{t+1}^{many > \beta'} \cap M} w_{m^{t+1}} \sum_{h \in \mathcal{H}} q_{t+1}(h) q_{t+1}(m^{t+1}|h) \\
= \; & \frac{2.3}{|\mathcal{H}|} \cdot \sum_{m^{t+1} \in M_{t+1}^{many > \beta'} \cap M} q_{t+1}(m^{t+1}) w_{m^{t+1}}
\end{aligned}$$

Let us focus on the sum over $h \in T_{m^{t+1}}$. From Claim 23 we know that

$$q_{t+1}(h|m^{t+1}) \leq 2.2 q_t(h|\psi(m^{t+1})) \quad (\star)$$

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 $\forall_i w_i \in [0,1]$)

$$\begin{aligned}
&\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta'}\cap M\\ h\in T_{m^{t+1}}}} q_{t+1}(h)q_{t+1}(h|m^{t+1})q_{t+1}(m^{t+1}|h)\\
&\qquad\leq \frac{4.4}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta'}\cap M\\ h\in T_{m^{t+1}}}} q_{t+1}(h)q_t(h|\psi(m^{t+1}))q_t(\psi(m^{t+1})|h)\\
\text{(see below)}\quad&\leq \frac{4.5}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta'}\cap M\\ h\in T_{m^{t+1}}\cap H_t}} q_t(h)q_t(h|\psi(m^{t+1}))q_t(\psi(m^{t+1})|h)\\
\text{(using Claim 22)}\quad&\leq \frac{4.5}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta'}\cap M\\ h\in T_{m^{t+1}}\cap H_t\\ m^t\in\psi(m^{t+1})}} q_t(h)q_t(h|m^t)q_t(m^t|h)\\
\text{(using Claim 12)}\quad&\leq \frac{4.5}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta'}\cap M\\ h\in T_{m^{t+1}}\\ m^t\in\psi(m^{t+1})}} q_t(m^t)q_t(h|m^t)^2\\
\text{(assumption in the claim)}\quad&\leq \frac{4.5}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta'}\cap M\\ h\in T_{m^{t+1}}\\ m^t\in\psi(m^{t+1})}} q_t(m^t)(a\cdot cer^t(M_t))^2\\
(|T_{m^{t+1}}|\leq\alpha'|\mathcal{H}|)\quad&\leq \frac{4.5}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta'}\cap M\\ m^t\in\psi(m^{t+1})}} q_t(m^t)(a\cdot cer^t(M_t))^2\cdot\alpha'|\mathcal{H}|\\
(cer^t(M_t)\leq\frac{3}{|\mathcal{H}|})\quad&\leq 13.5\alpha' a^2 cer^t(M_t)\cdot\frac{1}{|\mathcal{X}|}\sum_{\substack{m^{t+1}\in M_{t+1}^{many>\beta'}\cap M\\ m^t\in\psi(m^{t+1})}} q_t(m^t)\\
\text{(see below)}\quad&\leq 13.5\alpha' a^2 cer^t(M_t)
\end{aligned}$$

to understand why the second inequality is true, notice that we have a sum of the form

$$4.4\sum_{h\in T} q_{t+1}(h)a_h$$

for some value $a_h \geq 0$, which is equal (by the definition of q_t) to

$$\begin{aligned}
\frac{4.4}{|H_{t+1}|}\sum_{h\in H_{t+1}\cap T} a_h &\leq \frac{4.4}{|H_t|(1-1/c)}\sum_{h\in H_{t+1}\cap T} a_h \\
(\text{for } c \geq 45) &\leq \frac{4.5}{|H_t|}\sum_{h\in H_{t+1}\cap T} a_h \\
(H_{t+1}\subseteq H_t) &\leq \frac{4.5}{|H_t|}\sum_{h\in H_t\cap T} a_h \\
&= 4.5\sum_{h\in T} q_t(h)a_h
\end{aligned}$$

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}. ■

11. 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 $cer^1(M_1) = O\left(\frac{1}{|\mathcal{H}|}\right)$, 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 $O(1/|\mathcal{H}|)$ 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 20, 21), or is part of a *small* error set. To achieve this we duplicate some of the memories.

Claim 27 (Decomposition lemma) *Suppose that the hypotheses graph is an $(\alpha,\beta,\epsilon)-$H-expander, the number of memory states is at most Λ, $\sqrt{\alpha|\mathcal{H}|c\cdot cer^t(M_t)}+\epsilon < 1/44$, 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 $M_{t+1}^{many>\gamma_1}\cup M_{t+1}^{heavy>\gamma_2}$.

Recall the connection between q_t and G'_t mentioned in Section 5 — 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.

Proof For each hypothesis there are exactly $\frac{|\mathcal{X}|}{2}$ labeled examples that are consistent with it. Thus there are $\left(\frac{|\mathcal{X}|}{2}\right)^t$ sequences of length t that are consistent with each hypothesis. Thus, there are

$$L_t := \left(\frac{|\mathcal{X}|}{2}\right)^t |\mathcal{H}|$$

edges in the knowledge graph at time t. Put it in other words, at each time step the number of edges is multiplied by $\frac{|\mathcal{X}|}{2}$, i.e.,

$$L_{t+1} = \frac{|\mathcal{X}|}{2} \cdot L_t.$$

This implies that there are at least $(1-\gamma)L_{t+1} = (1-\gamma)\frac{|\mathcal{X}|}{2} \cdot L_t$ edges in G'_{t+1}.

We start by removing a small number of edges:

- Remove edges in G_{t+1} connected to memories m at time t with $\sum_h q_t(h|m)^2 > c \cdot cer^t(M_t)$
  - As was discussed in Section 6, thus are Bad_M and Markov's inequality implies that at most $1/c$ of the weight over memories are of this type. I.e., we remove at most $1/c$ fraction of the edges.
- Remove edges in G_{t+1} that the $t+1$ labeled example is a non-representative labeled example (see Section 7) in G_t (i.e., labeled examples x such that $q_{t+1}(x|m) \leq \frac{1}{1.1|\mathcal{X}|}$ or $q_{t+1}(x|m) \geq \frac{1.1}{|\mathcal{X}|}$)
  - Using Claim 18 for any memory m at time t, $|NRep(m)| \leq 2\beta|\mathcal{X}|$. Thus, the fraction of edges in G_{t+1} of this type is at most 4β (because for each memory m at least $\frac{|\mathcal{X}|}{2}$ of the possible $\mathcal{X}$ labeled examples will make a new edge in G_{t+1} and at most $4\beta\frac{|\mathcal{X}|}{2}$ of these edges are non-representative labeled examples).
- Remove edges in G_{t+1} connected to memories m^{t+1} at time $t+1$ with less than $(1-\gamma)L_{t+1}/(c\Lambda)$ edges (i.e., much less than the average number of edges to a memory)
  - A simple calculation proves that we removed at most $1/c$ of the edges.
- Using the remaining edges, for each other memory m^{t+1} create a few copies in the following way (note that we will not add new edges in the process):

many-source Do until impossible: if m^{t+1} is connected to memories $\{m_i^t\}$ with total q_t-weight more than γ_1 using labeled examples $\{S_i\}$, then create a copy (m^{t+1}, j) and connect all memories m_i^t to (m^{t+1}, j) with one labeled example $x_i \in S_i$. Retain all other labeled examples $S_i \setminus \{x_i\}$.

heavy-source Do until impossible: if m^{t+1} is connected to a memory m^t with more than $\gamma_2|\mathcal{X}|$ labeled examples S, then create a copy (m^{t+1}, j) and connect memory m^t to (m^{t+1}, j) with all labeled examples S.

- If some edges are still connected to m^{t+1} after the last two steps are over, then remove those remaining edges.
- Let us explore how many edges were removed in this step. For that, first recall what γ_1 and γ_2 represent (see Definitions 20, 21). If there are more than $\gamma_1|\mathcal{X}|\gamma_2 L_t$ edges connected to a memory m^{t+1} in graph G'_{t+1}, then we know that the there is i such hat memory (m^{t+1}, i) is heavy-sourced or many-sourced.

  What is the fraction of edges we might remove out of all edges entering a memory at time $t+1$? (recall that memories left have at least than $(1-\gamma)L_{t+1}/(c\Lambda)$ edges entering them)

$$\frac{\gamma_1|\mathcal{X}|\gamma_2 L_t}{(1-\gamma)L_{t+1}/(c\Lambda)} = \frac{\gamma_1|\mathcal{X}|\gamma_2 L_t}{(1-\gamma)L_t|\mathcal{X}|/(2c\Lambda)} = 2c\gamma_1\gamma_2\Lambda \cdot \frac{1}{1-\gamma} \leq 4c\gamma_1\gamma_2\Lambda,$$

  where in the last inequality we used $\frac{1}{1-\gamma} \leq 2$, which is true since $\gamma \leq 0.5$.

Thus, the total fraction of edges removed in the entire process is at most

$$\frac{1}{c} + 4\beta + \frac{1}{c} + 4c\gamma_1\gamma_2\Lambda.$$

■

Recall that we defined M_1 in Definition 16 and before that we defined ϵ_1. For all $t \geq 1$, we will construct M_{t+1} formally in the proof of Claim 28. Recall also that H_{t+1} and c were defined in Section 6.

It might be helpful to think of d in the following claim as roughly $\sqrt{|\mathcal{H}|}$, $c = |\mathcal{H}|^s$, for a very small constant $s > 0$, and $|\mathcal{H}| \approx |\mathcal{X}|$.

Claim 28 *If the hypotheses graph is d-mixing, $\beta = \frac{c^{100}\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}$, Λ is the number of memory states with $\Lambda \leq (c\beta)^{-1.25}$, and $c > 10^8$, 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 an $(\frac{1}{c^{10}}, 2\beta c^{16}, \frac{t+1}{c^{14}})-$ 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, $M \subseteq M_t$*

$$cer^t_w(M) \leq \left[\frac{2.3}{|\mathcal{H}|}\left(\sum_{m\in M} q_t(m)w_m\right) + \frac{8}{c}\cdot\sum_{t'=1}^{t-1} cer^{t'}(M_{t'})\right]\left(1+\frac{6}{c}\right)^{t-1}$$

- $q_t(M_t) \geq 1 - \frac{2t}{c}$
- *for each $h \in H_t, m \in M_t$ 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 29) that the last item in the claim's list implies that $cer^t(M_t) \leq \frac{3}{|\mathcal{H}|}$.

Claim 29 *If* $cer^1(M_1) \leq \frac{2.4}{|\mathcal{H}|}$ *and for any* $t \leq 10^{-8} \cdot c$,

$$cer^t(M_t) \leq \left[\frac{2.3}{|\mathcal{H}|} + \frac{8}{c} \cdot \sum_{t'=1}^{t-1} cer^{t'}(M_{t'})\right] \left(1 + \frac{6}{c}\right)^{t-1}$$

then $cer^t(M_t) \leq \frac{3}{|\mathcal{H}|}$.

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,

$$cer^t(M_t) \leq \frac{2.4}{|\mathcal{H}|} + \frac{8.5}{c} \cdot \sum_{t'=1}^{t-1} cer^{t'}(M_{t'}).$$

Let us focus on the following recursively defined series: $a_1 = \frac{2.4}{|\mathcal{H}|}$ and

$$a_{t+1} = \frac{2.4}{|\mathcal{H}|} + \frac{8.5}{c} \cdot \sum_{t'=1}^{t} a_{t'}.$$

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

$$\begin{aligned}
a_{t+1} &\leq a_1 + \frac{8.5t}{c} a_t \\
(t \leq 10^{-8} \cdot c) &\leq a_1 + \frac{1}{100} a_t \\
&\leq a_1 + \frac{1}{100}(a_1 + \frac{1}{100} a_{t-1}) \\
\text{(geometric series)} &\leq \ldots \leq 1.02 a_1 \leq \frac{3}{|\mathcal{H}|}
\end{aligned}$$

■

Proof (of Claim 28) From Proposition 9 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 11, it is also $(\alpha, \beta, \epsilon_1) -$ H-expander with $\epsilon_1 = \frac{2\mathrm{d}}{\sqrt{\alpha|\mathcal{H}|\beta|\mathcal{X}|}}$ for any α. We pick $\alpha = 1/c^{34}$. By the choice of α, β and for $c \geq 2$ we have that

$$\epsilon_1 = \frac{2\mathrm{d}}{\sqrt{\alpha|\mathcal{H}| \cdot \frac{c^{100}\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|} \cdot |\mathcal{X}|}} = \frac{2}{\sqrt{\alpha c^{100}}} \leq \frac{1}{c^{17}}.$$

Those values of α, β will be our choice for α_1 and β_1 that appear before Definition 16.

We prove the claim by induction on t.

Induction Basis. At the beginning $H_1 = \mathcal{H}$. From Claim 10, we know that G'_1 is an $(\alpha, \beta, \epsilon = 8\epsilon_1 + \alpha) -$ K-expander. Note that $\epsilon \leq 1/c^{16}$ for $c \geq 10$.

Denote by E is the set of edges in the hypotheses graph.

Take memory m in M_1, and denote its degree by d_m, then m's certainty is equal to

$$\sum_h q_1(h|m)^2 = \sum_h \left(\frac{I_{(h,m)\in E}}{d_m}\right)^2 = d_m \left(\frac{1}{d_m}\right)^2 = \frac{1}{d_m},$$

where the second equality follows from the fact that for each $(h, m) \notin E$ the value in the sum is equal to 0 and for each $(h, m) \in E$, the value in the sum is equal to $1/d_m^2$, and there are exactly d_m hypotheses h with $(h, m) \in E$. From the definition of M_1 (see Definition 16), using $\epsilon_1 \leq 0.04$ we know that

$$d_m \geq |\mathcal{H}|(1/2 - \epsilon_1) \geq |\mathcal{H}|/2.2,$$

hence

$$cer^1(M_1) \leq \frac{2.2}{|\mathcal{H}|}$$

We remove all edges touching a memory not in M_1. From Claim 7 there are at most $2\beta|\mathcal{X}|$ memories not in M_1. The number of edges connected to each memory is at most $|\mathcal{H}|$. I.e., we remove at most $2\beta|\mathcal{X}||\mathcal{H}|$ edges out of the $|\mathcal{X}||\mathcal{H}|/2$. In other words, we remove at most 4β fraction of the edges.

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 14 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 $t - 1 \leq 0.5 \cdot c \leq \frac{c \ln 3}{2}$.

We will use Claim 19 to prove the K-expander property of G'_{t+1}. Note that for $c \geq 48$,

$$16\epsilon + 16\sqrt{3\alpha c} + \frac{2\beta}{2\beta c^{16}} \leq \frac{1}{c^{15}}$$

Thus, using Claim 19 and the inductive hypothesis, the graph G'_{t+1} is a $(\frac{1}{c^{10}}, 2\beta c^{16}, \frac{t+2}{c^{14}})-$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. Note also that for $c \geq 2$ it holds that $\sqrt{3\alpha c} + \epsilon < 1/44$.

We use Claim 27 with

- Let γ_1 define the many-source set $M_{t+1}^{many>\gamma_1}$ (see Definition 21). To later apply Claim 26, we choose $\gamma_1 = 2\beta c^{16}$.
- Let γ_2 define the heavy-source set $M_{t+1}^{heavy>\gamma_2}$ (see Definition 20). To later apply Claim 21 we choose

  $$\gamma_2 = 5\epsilon_{sam} c + 50c^5 \sqrt{\frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|\epsilon_{sam}^2}}$$

  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 $\epsilon_{sam}^2 = 10c^4 \sqrt[2]{\frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}}$, which means that $\gamma_2 < 100c^3 \sqrt[4]{\frac{\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}}$.

For later use, notice that

$$\gamma_1\gamma_2 \leq 200\beta c^{19}\sqrt[4]{\frac{\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}} = 200\beta c^{19}\sqrt[4]{\frac{\beta}{c^{100}}} = \frac{200}{c^6}\beta^{1.25}$$

From Claim 27 we know that by removing at most

$$\frac{2}{c} + 4\beta + \frac{800}{c^5}\beta^{1.25}\Lambda \leq \frac{2}{c} + 4\beta + \frac{\beta^{1.25}}{c}\Lambda$$

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 21 to prove that

$$cer_w^{t+1}(M_{t+1}^{heavy>\gamma_2} \cap M)$$

is smaller than

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

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 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 16

$$\begin{aligned}
q_{t+1}(S|m^t) &= \sum_h \Pr(S|h) q_t(h|m^t) \\
&\geq \sum_{h \notin Err(S)} \Pr(S|h) q_t(h|m^t) \\
&\geq \sum_{h \notin Err(S)} \left(\frac{|S|}{|\mathcal{X}|} - \epsilon_{sam} \right) q_t(h|m^t) \\
&= \frac{|S|}{|\mathcal{X}|} \left(1 - \frac{\epsilon_{sam}}{|S|/|\mathcal{X}|} \right) \sum_{h \notin Err(S)} q_t(h|m^t) \\
(\text{definition of } \gamma_2) \quad &\geq \frac{|S|}{|\mathcal{X}|} \left(1 - \frac{\epsilon_{sam}}{5\epsilon_{sam} c} \right) \sum_{h \notin Err(S)} q_t(h|m^t) \\
(\text{Claim 1 } \& cer^t(M_t) \leq \frac{3}{|\mathcal{H}|}) \quad &\geq \frac{|S|}{|\mathcal{X}|} \left(1 - \frac{1}{5c} \right) (1 - \sqrt{3c\epsilon'_{sam}})
\end{aligned}$$

This means that

$$\frac{|S|}{|\mathcal{X}|} \leq \frac{q_{t+1}(S|m^t)}{\left(1 - \frac{1}{5c}\right)\left(1 - \sqrt{3c\epsilon'_{sam}}\right)}.$$

So we just need to show that

$$\frac{1}{\left(1 - \frac{1}{5c}\right)\left(1 - \sqrt{3c\epsilon'_{sam}}\right)} \leq 1 + \frac{1}{c}$$

First let us simplify $\sqrt{3c\epsilon'_{sam}}$

$$\begin{aligned}
\sqrt{3c\epsilon'_{sam}} &= \sqrt{3c \frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|\epsilon^2_{sam}}} \\
\left(\frac{\beta}{c^{100}} = \frac{\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|} \right) &= \sqrt{3c \frac{8\beta}{c^{100} 10 c^4 \sqrt{\frac{8\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}}}} \\
&= \sqrt{3c \frac{8\beta}{c^{100} 10 c^4 \sqrt{\frac{8\beta}{c^{100}}}}} \\
(\beta \leq 1) &= \frac{1}{c^{26.5}} \sqrt{\frac{3\sqrt{8}}{10}} \\
&\leq \frac{1}{4c}
\end{aligned}$$

Note that

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

Many-sourced memories. We can use Claim 26 since $\frac{1}{c^{10}} \geq 22 \cdot \frac{t+1}{c^{14}}$ and we get that

$$\begin{aligned}cer^{t+1}(M_{t+1}^{many>\gamma_1} \cap M) &\leq \frac{2.3}{|\mathcal{H}|} \cdot \sum_{m \in M_{t+1}^{many>\gamma_1} \cap M} q_{t+1}(m) w_m + \frac{1}{c} \cdot cer^t(M_t)\\ &= \left[\frac{2.3}{|\mathcal{H}|} \cdot \sum_{m=\{(m_i^t,x_i)\} \in M_{t+1}^{many>\gamma_1} \cap M} q_t(m_i^t) q_{t+1}(x_i|m_i^t) w_m\right] +\\ &+ \frac{1}{c} \cdot cer^t(M_t) \quad (\star\star)\end{aligned}$$

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

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

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

$$\begin{aligned}cer_w^t(M) &\leq \sum_{m^t} q_t(m^t) \left(\sum_h q_t^2(h|m^t) \cdot w_{m^t}^{heavy} + \frac{2.3}{|\mathcal{H}|} \cdot w_{m^t}^{many}\right) + \frac{8}{c} \cdot cer^t(M_t)\\ &\leq \sum_{m^t} q_t(m^t) \max\left\{\sum_h q_t^2(h|m^t), \frac{2.3}{|\mathcal{H}|}\right\} \cdot w_{m^t} + \frac{8}{c} \cdot cer^t(M_t)\end{aligned}$$

Define $M_a = \{m^t | \sum_h q_t^2(h|m^t) > 2.3/|\mathcal{H}|\}$ and $M_b = \{m^t | \sum_h q_t^2(h|m^t) \leq 2.3/|\mathcal{H}|\}$ and the last term is equal to

$$\left[\sum_{m^t \in M_a} q_t(m^t) w_{m^t} \cdot \sum_h q_t^2(h|m^t)\right] + \left[\sum_{m^t \in M_b} q_t(m^t) w_{m^t} \cdot \frac{2.3}{|\mathcal{H}|}\right] + \left[\frac{8}{c} \cdot cer^t(M_t)\right]$$

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

$$\left[\frac{2.3}{|\mathcal{H}|}\left(\sum_{m^t\in M_a} q_t(m^t)w_{m^t}\right)+\frac{8}{c}\cdot\sum_{t'=1}^{t-1}cer^{t'}(M_{t'})\right]\left(1+\frac{6}{c}\right)^{t-1}+$$
$$+\quad\left[\sum_{m^t\in M_b} q_t(m^t)w_{m^t}\cdot\frac{2.3}{|\mathcal{H}|}\right]+\left[\frac{8}{c}\cdot cer^t(M_t)\right]$$

which is at most

$$\left[\frac{2.3}{|\mathcal{H}|}\left(\sum_{m^t\in M_t} q_t(m^t)w_{m^t}\right)+\frac{8}{c}\cdot\sum_{t'=1}^{t}cer^{t'}(M_{t'})\right]\left(1+\frac{6}{c}\right)^{t-1}$$

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

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

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

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

recall that M' contains all the memories that are heavy-sourced or many-sourced. Thus, using also Claim 11, 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 27. Thus (using the bound we showed earlier on $\gamma_1\gamma_2$), in time $t+1$ we removed a total fraction of

$$\left(\frac{1}{c} + \frac{1}{c^2}\right) + \left(\frac{2}{c} + 4\beta + \frac{\beta^{1.25}}{c}\Lambda\right)$$

edges. We will prove that this term is at most $\frac{4}{c}$. From the assumption in the claim we know that $\Lambda \leq (c\beta)^{-1.25}$, this means that $\Lambda\beta^{1.25} \leq \frac{1}{c^{1.25}} \leq \frac{1}{2c}$. Also, for the claim to be nontrivial, $\Lambda \geq |\mathcal{H}|$, thus, the term $(c\beta)^{1.25}$ must be smaller than $1/|\mathcal{H}|$. In particular, $16\beta c \leq 1$. Hence, the total fraction of edges removed at time $t+1$ is at most

$$\frac{1}{c} + \frac{1}{16c} + \frac{2}{c} + \frac{1}{4c} + \frac{1}{2c} \leq \frac{4}{c}.$$

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,

$$cer_w^{t+1}(M) \leq \left[\frac{2.3}{|\mathcal{H}|}\left(\sum_{m\in M} q_t(m)w_m\right) + \frac{8}{c}\cdot\sum_{t'=1}^{t} cer^{t'}(M_{t'})\right]\left(1+\frac{6}{c}\right)^t$$

■

11.1. Choosing Parameters

For a mixing hypothesis class $\mathcal{H}$, i.e., when $d^2 \approx |\mathcal{X}|$, we show a lower bound that is roughly $|\mathcal{H}|^{1.25}$ on the number of memory states needed for learning $\mathcal{H}$ using less than $|\mathcal{H}|^{\Theta(1)}$ labeled examples.

Theorem 23 *For any constant $s \in (0,1)$, if the hypotheses graph is d-mixing, $|\mathcal{H}|$ is at least some constant, and the number of memory states is bounded by*

$$\left(\frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}\right)^{1.25} \cdot \frac{1}{\left(1+\frac{16\mathrm{d}^2}{|\mathcal{X}|}\right)^{1.25}|\mathcal{H}|^s}$$

then any learning algorithm that outputs the underlying hypothesis (or an approximation of it) with probability at least $1/3$ must use at least $10^{-8}|\mathcal{H}|^{s/130}$ examples.

Proof From Claim 9, we know that there is an hypothesis class $\mathcal{H}' \subseteq \mathcal{H}$ with $|\mathcal{H}'| \geq \frac{|\mathcal{H}|}{1+\frac{16\mathrm{d}^2}{|\mathcal{X}|}}$ such that every two hypotheses in $\mathcal{H}'$ has agreement less than $3/4$.

To apply Claim 28 to $\mathcal{H}'$ we will prove that the number of memory states $\Lambda \leq (c\beta)^{-1.25}$ with

$$\beta = \frac{c^{100}\mathrm{d}^2}{|\mathcal{H}'||\mathcal{X}|} \leq \frac{\left(1+\frac{16\mathrm{d}^2}{|\mathcal{X}|}\right)c^{100}\mathrm{d}^2}{|\mathcal{H}||\mathcal{X}|}.$$

Thus,

$$\frac{1}{(\beta c)^{1.25}} \geq \left(\frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}\right)^{1.25} \cdot \frac{1}{c^{130}\left(1+\frac{16\mathrm{d}^2}{|\mathcal{X}|}\right)^{1.25}}$$

Hence, using the assumption in the claim with $c^{130} = |\mathcal{H}|^s$, we have that $\Lambda \leq (c\beta)^{-1.25}$. From Claim 28 we can deduce that even after $10^{-8} \cdot c$ examples given, the certainty is at most $3/|\mathcal{H}'|$, the total number of edges removed is at most $\frac{4t}{c}$, and $1 - q_t(M_t) \leq \frac{2t}{c}$.

Using Claim 15 there is an hypothesis $h \in \mathcal{H}'$ such that the probability to correctly return it is at most

$$3\sqrt{c \cdot \frac{3}{|\mathcal{H}'|}} + 3 \cdot \frac{2t}{c} + \frac{4t}{c}$$

we will prove that this expression is at most $1/3$. Since the claim is nontrivial $\left(\frac{|\mathcal{H}||\mathcal{X}|}{\mathrm{d}^2}\right)^{1.25} > |\mathcal{H}|$. This implies that $\mathcal{H}'$ is much bigger than c. Hence, the first term that comprise this expression is much smaller than $1/3$. The sum of the last two terms that comprise this expression is much smaller than $1/3$ since they are equal to $7 \cdot 10^{-8}$.

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. ■

12. Applications

12.1. Random Hypothesis Class

One immediate consequence of Theorem 23, that uses the fact that random graphs are mixing (see e.g.,Krivelevich and Sudakov (2006)), is that almost all hypothesis classes are unlearnable with bounded memory. Note that unlike for circuits, such a result does not follow from counting arguments because the number of possible hypotheses classes is $\binom{2^{|\mathcal{X}|/2}}{|\mathcal{H}|} \leq 2^{|\mathcal{X}||\mathcal{H}|}$, whereas the number of learners with Λ memory states and T labeled examples is about $\Lambda^{\Lambda T|\mathcal{X}|}$. For parameters of interest, like $\Lambda = |\mathcal{H}|^{\Theta(1)}$ and $T = |\mathcal{H}|^{\Theta(1)}$, the number of learners is much larger than the number of hypotheses classes.

Theorem 24 *A random hypothesis class with n hypotheses and n labeled examples (for sufficiently large n) almost surely cannot be learned with bounded memory, i.e., for any constant $0 \leq s \leq 1$ if the number of memory states is bounded by*

$$n^{1.25-s}$$

then any learning algorithm that outputs the underlying hypothesis (or an approximation of it) with probability at least $1/3$ must use at least $|\mathcal{H}|^{s/129}$ examples.

Proof A random class can be viewed as a random bipartite graph in the following way: the vertices are examples and hypotheses and there is an edge (h, x) between hypothesis h and example x if and only if $h(x) = 1$. We know that this graph with $|\mathcal{H}| = |\mathcal{X}| = n$ is almost surely $O(\sqrt{n})$-mixing and has degree $(1 + o(1))n/2$ (see Krivelevich and Sudakov (2006)). From Claim 5 we also know that the hypotheses graph is also $O(\sqrt{n})$-mixing. Now we can apply Theorem 23 to deduce the unlearnability of a random hypothesis class. ■

12.2. Error Correcting Codes

The next claim is helpful in proving that an hypothesis class is mixing:

Theorem 25 (see Thomason (1989)) *Let (A, B, E) be a bipartite graph with $|A| = |B| = n$. Let each vertex in A have degree at least pn, where $1/2 \leq p < 1$, and let $\mu \geq 0$ be such that no two vertices of A have more than $p^2 n + \mu$ common neighbors. Then, (A, B, E) is $\sqrt{(p + \mu) \cdot n}$-mixing.*

We consider one use of this theorem, by proving that any hypothesis class that is also an *error correcting code* (will be defined formally) cannot be learned with bounded memory.

Definition 26 (Code) *A binary* code *is a subset $C \subseteq \{0, 1\}^n$. The elements of C are called the* codewords *in C.*

Definition 27 (Distance) *An error correcting code $C \subseteq \{0, 1\}^n$ has (relative) distance δ if for any $c_1 \neq c_2 \in C$, the fraction of coordinates that c_1 and c_2 differ is at least δ.*

A code $C \subseteq \{0,1\}^n$ can be viewed as an hypothesis class $\mathcal{H}_C$: the hypotheses correspond to the codewords, the examples correspond to the n coordinates, and an hypothesis $h_c \in \mathcal{H}_C$ is defined by $h_c(i)$ which is equal to the i-th coordinate of c. So the number of labeled examples $\mathcal{X}$, is $|\mathcal{X}| = 2n$.

If a code C has distance at least $\delta \geq \frac{1}{2} - \epsilon$, then the number of common neighbors of any two hypotheses is at most

$$(1-\delta)n = (1-\delta)\frac{|\mathcal{X}|}{2} \leq \left(\frac{1}{2}+\epsilon\right)\frac{|\mathcal{X}|}{2} = \left(\frac{1}{2}\right)^2 |\mathcal{X}| + \epsilon\frac{|\mathcal{X}|}{2}$$

Denote $\mu = \epsilon\frac{|\mathcal{X}|}{2}$. To use Theorem 25 we need to make sure that $|C| \geq 2n$ (and take only $2n$ codewords from C as hypotheses) and then we can bound the mixing parameter by

$$\sqrt{\left(\frac{1}{2}+\mu\right)\frac{|\mathcal{X}|}{2}} = \sqrt{|\mathcal{X}|} \cdot \sqrt{\frac{1}{4} + \frac{\epsilon|\mathcal{X}|}{4}}$$

Summing up the discussion so far, using Theorem 23 with the mixing parameter $\mathrm{d}^2 = \frac{|\mathcal{X}|}{4}(1+\epsilon|\mathcal{X}|)$ and $|\mathcal{H}| = |\mathcal{X}| = 2n$, we have the following theorem.

Theorem 28 *For any code $C \subseteq \{0,1\}^n$ with $|C| = 2n$ and relative distance at least $\frac{1}{2} - \epsilon$ and any constant $s \in (0,1)$, if the number of memory states is bounded by*

$$\left(\frac{4|C|}{1+2\epsilon n}\right)^{1.25} \cdot \frac{1}{(5+8\epsilon n)^{1.25}\,|C|^s},$$

then any learning algorithm for $\mathcal{H}_C$ that outputs the underlying hypothesis (or an approximation of it) with probability at least $1/3$ must use at least $|C|^{s/130}$ examples.

Note that the theorem is useful for codes that have very small rate but very high distance (for reference, see, e.g., MacWilliams and Sloane (1977)).

References

M. Aigner. Turan's graph theorem. *American Mathematical Monthly*, 102(9):808–816, 1995.

Y. Bilu and N. Linial. Lifts, discrepancy and nearly optimal spectral gap. *Combinatorica*, 26(5): 495–519, 2006.

S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

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

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

F.J. MacWilliams and N.J.A Sloane. *The theory of error correcting codes*. Elsevier, 1977.

D. Moshkovitz and M. Moshkovitz. Mixing implies strong lower bounds for space bounded learning. Manuscript, 2017.

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.

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

O. Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 163–171, 2014.

J. Steinhardt, G. Valiant, and S. Wager. Memory, communication, and statistical queries. In *Conference on Learning Theory (COLT)*, 2016.

A. Thomason. Dense expanders and pseudo-random bipartite graphs. *Discrete Mathematics*, 75 (1-3):381–386, 1989.

LG. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.