Machine Learning, 9, 205-260 (1992)

On the Computational Complexity of Approximating Distributions by Probabilistic Automata

NAOKI ABE
Information Basic Research Laboratory, C&C Information Technology Research Laboratories, NEC Corporation, 4-1-1 Miyazaki, Miyamae-ku, Kawasaki 216 Japan

MANFRED K. WARMUTH
Computer Engineering and Information Sciences, University of California, Santa Cruz, CA 95064

Abstract. We introduce a rigorous performance criterion for training algorithms for probabilistic automata (PAs) and hidden Markov models (HMMs), used extensively for speech recognition, and analyze the complexity of the training problem as a computational problem. The PA training problem is the problem of approximating an arbitrary, unknown source distribution by distributions generated by a PA. We investigate the following question about this important, well-studied problem: Does there exist an *efficient* training algorithm such that the trained PAs *provably converge* to a model close to an optimum one with high confidence, after only a feasibly small set of training data? We model this problem in the framework of computational learning theory and analyze the sample as well as computational complexity. We show that the number of examples required for training PAs is moderate—except for some log factors the number of examples is linear in the number of transition probabilities to be trained and a low-degree polynomial in the example length and parameters quantifying the accuracy and confidence. Computationally, however, training PAs is quite demanding: Fixed state size PAs are trainable in time polynomial in the accuracy and confidence parameters and example length, but *not* in the alphabet size unless **RP** = **NP**. The latter result is shown via a strong non-approximability result for the single string maximum likelihood model probem for 2-state PAs, which is of independent interest.

Keywords. Hidden Markov models, PAC learning model, density estimation, Kullback-Leibler divergence, computational learning theory

1. Introduction

We address the problem of approximating an arbitrary, unknown source distribution by distributions generated by probabilistic automata. Probabilistic automata (PAs), and hidden Markov models[1] (HMMs) which are closely related to PAs, are used extensively as models for probabilistic generation of speech signals for the purpose of speech recognition (see for example Levinson, Rabiner & Sondhi, (1983)). The problem addressed in the present paper corresponds to that of training a parameterized hidden Markov model for a particular spoken word with a set of actual speech signals for that word. In particular, we are interested in the question of whether there exists an algorithm that, when given a sample generated from an arbitrary unknown target distribution, outputs a probabilistic automaton that approximates the unknown distribution 'as closely as possible,' that is, with high probability the distribution induced by the output PA is sufficiently close to an 'optimal' one among all possible probabilistic automata satisfying a certain *prescribed constraint.* Here a *constraint* is given to the algorithm in the form of a subset of the state set specifying

the *legal initial states* and a labeled directed graph specifying the set of *legal transitions.* The training problem, therefore, is the problem of finding a near optimal setting of the initial and transition probabilities on the legal initial states and transitions in the input constraint. A *class* of constraints is said to be trainable with sample complexity $q(\ldots)$ if there exists an algorithm which trains every constraint in the class, and the sample size required for a given acuracy, confidence, length of the example strings, and the size of the input constraint is bounded above by the function q of these parameters. Here the size of the input constraint translates to the number of probability parameters being trained. A class of constraints is said to be *polynomially trainable* if there is a training algorithm with polynomial sample complexity whose running time is polynomial in the total sample length. Of particular interest to us is the special case of this problem in which the input constraint is null, namely all initial states and transitions are legal. This special case translates to the problem of finding a near optimal probabilistic automaton with a *given number of states.*

Our model is a natural adaptation of the PAC-learning paradigm of Valiant (1984) and Blumer, et al. (1989) and is inspired by the model of efficient unsupervised learning of Laird (1988). It is also related to the models for learning languages from stochastic data in the limit proposed and studied by Angluin (1988). Our formulation requires the algorithm to be particularly *robust* in the sense that we do not assume anything about the target distribution—a formulation which is closely related to the 'robust' generalization of the PAC paradigm proposed by Haussler (1991). The distance measure between the distributions used in this paper to evaluate the accuracy of a hypothesis with respect to the target distribution is the well-known 'Kullback-Leibler divergence' (Kullback, 1967). Other commonly used measures of distance between probability distributions are, for example, the χ^2 distance, the variation distance, the quadratic distance (Kearns & Schapire, 1990), and the Hellinger distance (Barron & Cover, 1989). The Kullback-Leibler divergence is a standard notion of distance, which enjoys many desirable properties (see Section 2). Furthermore, the Kullback-Leibler divergence is known to bound from above the Hellinger distance as well as half the square of the variation distance and of the quadratic distance. These relationships for the more general case of *conditional distributions* are surveyed by Yamanishi (1991).

Using this model, we give a number of results: We show that an arbitrary class of constraints is trainable by exhibiting a training algorithm whose sample comlexity is essentially linear in the size of the constraint being trained and a low-degree polynomial in the example length and parameters quantifying the accuracy and confidence. In addition, the running time of our training algorithm is polynomial in the total sample length if the size of the input constraint is bounded by a constant, thus showing that finite classes are polynomially trainable. In particular, an arbitrary *fixed* constraint is trainable in time polynomial in the accuracy and confidence parameters and the example length. If the alphabet size is variable, however, no polynomial time training algorithm exists for the class of 2-state null constraints[2], unless **RP** = **NP**.

To the best of our knowledge, our upper bound is the first rigorous result on the sample complexity of training PAs and HMMs, with respect to the classical measure of Kullback-Leibler divergence. Our proof is also interesting in the sense that we manage to get around the problem caused by the fact that the Kullback-Leibler divergence is unbounded. This property prohibits the direct use of certain useful techniques such as Hoeffding's inequality

for showing uniform convergence of empirical estimates of random variables to their true means. We get around this problem roughly as follows: We bound the smallest transition probabilities in the training algorithm's hypotheses from below by a decreasing function of the sample size m. We show uniform convergence for these successive classes of 'bounded' probabilistic automata using Hoeffding's inequality. We then show that for sufficiently large m, an optimum automaton in the m-th bounded class is close to an optimum one in the entire class with high probability. Interestingly, the trick of bounding probabilities away from zero is often used in practice in an attempt to solve what is known as the 'finite sample problem' (Levinson, Rabiner & Sondhi, 1983). Our result provides a rigorous justification for a particular way of setting those probability bounds, from the point of view of proving bounds on the sample complexity.

The sample complexity bound we obtain allows us to extend the classical equivalence between the minimization of the Kullback-Leibler divergence with respect to the *empirical* distribution and the maximization of the likelihood of the given data: We show that the polynomial time trainability of a class of constraints C is equivalent to the polynomial time approximability of the 'maximum likelihood model' problem (MLM) for the same class C—the problem of setting the initial and transition probabilities in a given constraint in C so that the probability assigned on a given finite sample is maximized. More precisely, we show that the polynomial time trainability of a class of constraints C is equivalent to the approximability of the MLM problem for C with a factor $1 + \epsilon$ in random time polynomial in $1/\epsilon$ and the size t of the input constraint. Furthermore, we show that this latter notion of approximability of the MLM problem for C is also equivalent to a seemingly much weaker notion of approximability: Approximability within factor $2^{p(n,t)m^{\alpha}}$ in random polynomial time, where m is the sample size, α is an arbitrary constant less than 1, and $p(n,t)$ is a polynomial in the example length n and the size of the input constraint t. We use the above equivalence between the training and MLM problems to show our hardness result: For *variable* alphabet size, the MLM problem for 2-state *null* constaints, or the problem of finding a 2-state PA assigning the maximum likelihood on the input sample, is hard to approximate (unless **P** = **NP**), and hence the class of 2-state null constraints is not polynomially trainable (unless **RP** = **NP**).

The hardness result for the MLM problem for 2-state null constaints is shown via the following non-approximability result for the *single string* MLM problem for the same class—the special case of MLM in which the input sample consists of a single string. We show that it is hard to approximate the single string MLM problem for the 2-state null constraints within a factor of $2^{|w|^{1-\alpha}}$ for any positive constant α, where w is the input word, in time polynomial in the word length and alphabet size, unless **P** = **NP**. Note that it is a very strong non-approximability result[3], since there is a trivial training algorithm, using only 1-state probabilistic automata, that can guarantee approximation within a factor of $2^{|w|}$ of the best 2-state PA. The proof of the hardness result uses as a starting point the type of technique commonly used in the learning theory literature for showing the hardness of a 'sample consistency' or 'minimum consistent concept' problem in discrete domains such as automata and boolean formulas (Gold, 1978; Angluin, 1978; Pitt & Warmuth, 1989). In particular, our proof makes use of notions used in Angluin's proof of the NP-completeness of the sample consistency problem for 2-state DFA (Angluin, 1989). The proof given here is, however, significantly more complex than the proof of the discrete

case, since corresponding to 'consistency' we have 'probability,' which is continuous and is thus much harder to get a hold of. For example, in our reduction of the satisfiability problem to the MLM problem for the 2-state null constraints, it is already non-trivial to formalize how a truth assignment is is to be simulated by a PA. We let each truth assignment correspond *conceptually* to one of 2^n many *deterministic* PAs of a *particular* kind, and for all the other (infinitely many) PAs, we quantify how 'far' they are from those corresponding to truth assignments. We then show that any PA that assigns the input word w a probability at least $1/2^{|w|^{1-\alpha}}$ times the probability assigned on w by an optimum PA must be 'close' to a deterministic PA corresponding to a satisfying assignment.

This paper is outlined as follows. We begin in Section 2 with some preliminary definitions and give the proof of the sample size bounds for training PAs in Section 3. In Section 4 we show the equivalence between the training problem and the approximate MLM problem for any class of constraints. In Section 5 we give the hardness result for the single string MLM problem for 2-state null constraints. Parts of this lengthy proof are given in Appendices A and B. In Section 6 we discuss briefly how the results of this paper apply to HMMs. We conclude by discussing a number of open problems inspired by this research in Section 7.

2. Preliminaries

This paper deals with approximating a probability distribution over words over some finite alphabet, Σ. For simplicity, we assume that all words with positive probability have the same length, n, i.e. the distribution is over the domain Σ^n. We call an element of Σ^n an *example*. A *sample* Ξ of Σ^n is a finite sequence of examples of Σ^n, $\Xi = \langle w_1, \ldots, w_m \rangle$, where m is the sample size. We abuse notation and write $x \in \Xi$ to mean that x appears in the sequence Ξ. We let $\#(x, \Xi)$ denote the number of occurrences of example x in sample Ξ. Using the above notation, we define the notion of the empirical distribution of a sample.

Definition 2.1. *Given a sample Ξ of size m of Σ^n, the empirical distribution of Ξ over Σ^n, written $\hat{D}_\Xi$, is defined by:*

$$\forall x \in \Sigma^n \; \hat{D}_\Xi(x) = \frac{\#(x, \Xi)}{m}$$

Note that for any y not in Ξ, we have $\hat{D}_\Xi(y) = 0$.

The probabilistic automation is formalized as a stochastic matrix M together with an 'initial distribution' π over the set of states. Intuitively, the probabilistic automation is much like a non-deterministic finite state automaton except that the transitions take place with probabilities prescribed by M. (See Figure 1.) To start the process, the machine chooses the initial state according to the initial distribution π, and then at any given point after that, the machine is in some state i, and at the next time step moves to another state j outputting some letter z, with probability specified by $M(i, j, z)$. If one stops the machine

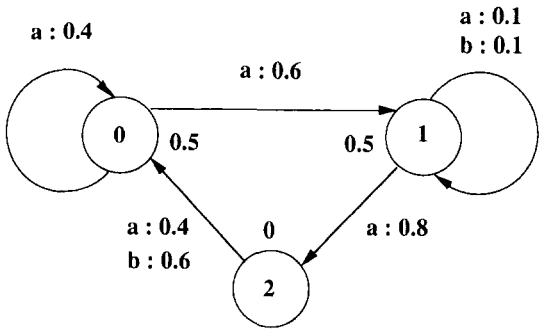

Figure 1. An example probabilistic automaton.

at time step n the machine ends in some state having generated a string of length n. In this way, a probabilistic automaton naturally defines a probability distribution over the set of strings of length n, for any particular n.

Defintion 2.2 (Probabilistic Automata (PA)). *A probabilistic automaton P is a quadruple $\langle S_P, \Sigma_P, \pi_P, M_P \rangle$ where S_P is a finite set of states, Σ_P is a finite alphabet, $\pi_P : S_P \to [0, 1]$ is a probability distribution over S_P, and $M_P : S_P \times S_P \times \Sigma_P \to [0, 1]$ is a stochastic matrix*[4]*, i.e.*

$$\sum_{i \in S_P} \pi_P(i) = 1 \text{ and } \forall i \in S_P \sum_{j \in S_P, z \in \Sigma_P} M_P(i, j, z) = 1 \tag{2.1}$$

Each $\pi_P(i)$ is called an initial probability, and each $M_P(i, j, z)$ is called a transition probability. For any string $w = w_1 \ldots w_n \in \Sigma_P^*$, the generation probability assigned on it by $P = \langle S_P, \Sigma_P, \pi_P, M_P \rangle$ is computed as follows.

$$P(w_1 \ldots w_n) = \sum_{\langle i_0, \ldots, i_n \rangle \in S_P^{n+1}} \pi_P(i_0) \cdot \prod_{j=0}^{n-1} M_P(i_j, i_{j+1}, w_{j+1}) \tag{2.2}$$

Thus, for any given example length n, P defines a probability distribution over Σ_P^n.

For example, the probability assigned by the probabilistic automaton P shown in Figure 1 on the string $w = aab$ is calculated as follows:

$$\begin{aligned} P(w) = {} & \pi_P(0) \cdot M_P(0, 0, a) \cdot M_P(0, 1, a) \cdot M_P(1, 1, b) \\ & + \pi_P(0) \cdot M_P(0, 1, a) \cdot M_P(1, 1, a) \cdot M_P(1, 1, b) \\ & + \pi_P(0) \cdot M_P(0, 1, a) \cdot M_P(1, 2, a) \cdot M_P(2, 0, b) \\ & + \pi_P(1) \cdot M_P(1, 1, a) \cdot M_P(1, 1, a) \cdot M_P(1, 1, b) \\ & + \pi_P(1) \cdot M_P(1, 1, a) \cdot M_P(1, 2, a) \cdot M_P(2, 0, b) \end{aligned}$$

$$= 0.5 \cdot 0.4 \cdot 0.6 \cdot 0.1 + 0.5 \cdot 0.6 \cdot 0.1 \cdot 0.1 + 0.5 \cdot 0.6 \cdot 0.8 \cdot 0.6 + 0.5 \cdot 0.1 \cdot 0.1 \cdot 0.1 + 0.5 \cdot 0.1 \cdot 0.8 \cdot 0.6$$
$$= 0.012 + 0.003 + 0.144 + 0.0005 + 0.024 = 0.1835$$

Note that for a probabilistic automaton P we use the same letter P to denote the probability distribution defined by P on Σ_P^n, where n will be clear from the context. A *PA constraint* is a quadruple $C = \langle \Sigma, S, I, G \rangle$ where *I is the inital state set* and G is the transition graph of C. I is a subset of the set S of all states. G is a subset of the set $S \times S \times \Sigma$ of all *transitions*. Note that a transition graph is a labeled directed graph with the state set S as the vertices, and the alphabet Σ as the set of labels. We write $|I|$ for the number of states in I, $|G|$ for the number of transitions in G. We then define the *size* of a constraint C, written $|C|$, as $|C| = |I| + |G|$. Note that the size of C corresponds to the number of probability parameters included in C.

We say that there is a path in C for a string $w_1 \ldots w_n \in \Sigma^n$, if there is a sequence of states $\langle i_0, \ldots, i_n \rangle$ from S^{n+1} such that $i_0 \in I$ and for all j, $1 \leq j \leq n$, $(i_{j-1}, i_j, w_j) \in G$. We say that a probabilistic automaton P *satisfies* a constraint $C = \langle \Sigma, S, I, G \rangle$, if and only if $S_P = S$ and $\Sigma_P = \Sigma$ and

$$\forall i \notin I,\ \pi_P(i) = 0 \text{ and}$$

$$\forall (i, j, z) \notin G,\ M_P(i, j, z) = 0 \tag{2.3}$$

Note that if a probabilistic automaton p satisfies a constraint C then for every word on which P assigns a positive probability, there is a path for it in C. If P satisfies C, one can think of π_P as a function from I into [0, 1], and M_P as a function from G into [0, 1]. We let $\mathcal{PA}(C)$ denote the class of probabilistic automata satisfying the constraint C. Note that $\mathcal{PA}(C) \subset [0, 1]^I \times [0, 1]^G$, where we let $[0, 1]^I$ denote the class of all functions from I into [0, 1], and $[0, 1]^G$ from G into [0, 1]. We also let $\mathcal{M}(G)$ denote the class of stochastic matrices M_P satisfying (2.3).

The notion of 'distance' among distributions we employ in this paper is a well-known measure in information theory called 'Kullback-Leibler divergence,' also known as the 'relative entropy.'

Definition 2.3 (Kullback-Leibler Divergence). *Let D and Q be probability distributions over countable domain X. The 'Kullback-Leibler divergence' of Q with respect to D, $d_{KL}(D, Q)$ is defined as follows.*

$$d_{KL}(D, Q) = \sum_{x \in X} D(x) \log \frac{D(x)}{Q(x)}$$

(Normally we think of D as the actual distribution against which a 'candidate' distribution Q is being compared. By convention, we let $0 \log 0 = 0$, *and* $0/0 = 1$.)

Note in the above definition that the base of logarithm is 2, following the conventions in coding theory. (Throughout this paper we let ln denote the natural logarithm and log the logarithm base 2.) The Kullback-Leibler divergence enjoys many natural properties: We can rewrite $d_{KL}(D, Q)$ as $E_D(\log 1/Q(\cdot)) - H(D)$ where $E_D(\log 1/Q(\cdot))$ is the expectation according to D of the random variable $\log 1/Q(\cdot)$, and $H(D)$ is the *entropy*[5] of D, defined as $\Sigma_{x \in X} D(x) \log 1/D(x)$. Recall that $\log 1/D(x)$ is the code length for x with respect to the "ideal code"[6] for D, and $H(D)$ is the expected code length of that code for the source distribution D (see Hamming (1986)). In other words, for the source distribution D, the divergence $d_{KL}(D, Q)$ measures the expected additional code length required when using the ideal code for Q instead of the ideal code for D. Thus, the ideal code of the distribution which minimizes the Kullback-Leibler divergence with respect to the source distribution also minimizes the expected code length of the future data. It is also well-known that minimizing the Kullback-Leibler divergence with respect to the *empirical* distribution $\hat{D}_{\Xi}$ observed in a sample $\Xi = \langle w_1, \ldots, w_m \rangle$ (Definition 2.1) corresponds to maximizing the likelihood of the sample, as demonstrated below. Minimizing $d_{KL}(\hat{D}_{\Xi}; Q)$ corresponds to minimizing $E_{\hat{D}_{\Xi}}(\log 1/Q(\cdot))$, and the following always holds:

$$E_{\hat{D}_{\Xi}}\left[\log \frac{1}{Q(\cdot)}\right] = \frac{1}{m} \cdot \sum_{i=1}^{m} \log \frac{1}{Q(w_i)} = \frac{1}{m} \log \prod_{i=1}^{m} \frac{1}{Q(w_i)} \tag{2.4}$$

Thus, $d_{KL}(\hat{D}_{\Xi}; Q)$ is minimized when $\Pi_{i=1}^{m} Q(w_i)$ is maximized, i.e. when Q maximizes the probability of having generated the sample. We summarize this as a lemma.

Lemma 2.1. *Let $\mathcal{P}$ be an arbitrary class of distributions over Σ^n, $\Xi = \langle w_1, \ldots, w_m \rangle$ an arbitrary sample of Σ^n, and $\hat{D}_{\Xi}$ the empirical distribution of Ξ. Then,*

$$E_{\hat{D}_{\Xi}}\left[\log \frac{1}{Q(\cdot)}\right] = \inf\{E_{\hat{D}_{\Xi}}\left[\log \frac{1}{P(\cdot)}\right] : P \in \mathcal{P}\}$$

if and only if

$$\prod_{i=1}^{m} Q(w_i) = \sup\{\prod_{i=1}^{m} P(w_i) : P \in \mathcal{P}\}$$

Below we give the definition of a training algorithm which is the central definition in this paper. Here we assume that a *randomized algorithm* has access to a fair coin and can flip it in a single time unit.

Definition 2.4 (Training PA Constraints). *A training algorithm takes as input a constraint C, a string length n, and a finite sample Ξ of Σ^n for some alphabet Σ, and outputs a probabilistic automaton which satisfies C. We say that a (possibly randomized) training algorithm A trains a class of constraints $\mathcal{C}$ with sample size $q(1/\epsilon, 1/\delta, n, t)$, if A, when given as*

input an arbitrary constraint $C \in \mathcal{C}$ of size t, a string length n, and a finite sample Ξ drawn independently at random from an arbitrary unknown distribution D over Σ^n, is such that whenever the sample size m exceeds $q(1/\epsilon, 1/\delta, n, t)$, then provided that $inf\{d_{KL}(D, P) : P \in \mathcal{PA}(C)\}$ is finite, A's output Q satisfies the following with probability at least $1 - \delta$:

$$d_{KL}(D, Q) - d_{KL}(D, Opt) \leq \epsilon$$

where Opt is a member of $\mathcal{PA}(C)$ satisfying:

$$d_{KL}(D, Opt) = min\{d_{KL}(D, P) : P \in \mathcal{PA}(C)\}$$

Here the probability is taken over the product distribution of D producing the sample and the random coin flips of A, if A is randomized. If there exists such a training algorithm then we say that $\mathcal{C}$ is trainable with sample complexity $q(1/\epsilon, 1/\delta, n, t)$. If there exists a training algorithm with a polynomial *sample complexity which also runs in time* polynomial *in the total sample length, then we say that $\mathcal{C}$ is polynomially trainable.*

Note that $inf\{d_{KL}(D, P) : P \in \mathcal{PA}(C)\}$ is infinite if and only if there is a word $x \in \Sigma^n$ such that $D(x) > 0$ and there is no path in C for x. In this case, $d_{KL}(D, P) = \infty$ for any P in $\mathcal{PA}(C)$. We still need to verify that *Opt* in the above definition is well-defined, when $inf\{d_{KL}(D, P) : P \in \mathcal{PA}(C)\}$ is finite. Define a function $\xi_D : \mathcal{PA}(C) \to [0, 1]$ for an arbitrary target distribution D by:

$$\xi_D(P) = \prod_{x \in \Sigma^n} P(x)^{D(x)}$$

Then since ξ_D is a continuous function, for an arbitrary D, mapping a compact subset of the parameter space $[0, 1]^I \times [0, 1]^G$, it attains a maximum at a particular PA, say *Opt'*:

$$\xi_D(Opt') = max\{\xi_D(P): P \in \mathcal{PA}(C)\}$$

Hence,

$$\begin{aligned} E_D\left[\log \frac{1}{Opt'(\cdot)}\right] &= \sum_{x \in \Sigma^n} D(x) \log \frac{1}{Opt'(x)} \\ &= \log \frac{1}{\xi_D(Opt')} \\ &= \log \frac{1}{max\{\xi_D(P) : P \in \mathcal{PA}(C)\}} \\ &= min\{\log \frac{1}{\xi_D(P)} : P \in \mathcal{PA}(C)\} \\ &= min\{E_D\left[\log \frac{1}{P(\cdot)}\right] : P \in \mathcal{PA}(C)\} \end{aligned} \tag{2.5}$$

Noting that $d_{KL}(D, P) = E_D(\log 1/P(\cdot)) - H(D)$, the equality (2.5) implies that $d_{KL}(D, Opt')$ equals $min\{d_{KL}(D, P) : P \in \mathcal{PA}(C)\}$. Now let $Opt = Opt'$ and we see that Opt is well-defined.

We define two versions of the maximum likelihood model problem for PAs considered in this paper.

Definition 2.5 Sample MLM Problem for C

Input: A constraint $C \in \mathcal{C}$, a string length n and a finite sample $\Xi = \langle w_1, \ldots, w_m \rangle$ of strings from Σ^n.

Output: A probabilistic automaton Q satisfying C which asigns the maximum genration probability (or the maximum likelihood) on Ξ among all such probabilistic automata, i.e.

$$\prod_{i=1}^{m} Q(w_i) = max \left\{ \prod_{i=1}^{m} P(w_i) : p \in \mathcal{PA}(C) \right\}$$

Again note that $max\{\Pi_{i=1}^{m} P(w_i) : P \in \mathcal{PA}(C)\}$ is well-defined because $\zeta : \mathcal{PA}(C) \to [0, 1]$ defined by

$$\zeta(P) = \prod_{i=1}^{m} P(w_i)$$

is a continuous function mapping a compact domain $\mathcal{PA}(C)$ into the range [0, 1].

The following definition is a special case of the sample maximum likelihood model problem in which the input sample consists of a single string. Note that for a single string, the initial probability distribution plays no significant role, because among probabilistic automata assigning the maximum probability on a given string there is always a probabilistic automaton in which exactly one state has initial probability one and other states have probability zero.

Definition 2.6 *For a stochastic matrix M and a word w, let $M(w)$ be the maximum generation probability assignable on w by M with the* best *initial state.*

Note that with this definition $M(uv) \leq M(u) \cdot M(v)$, in general. The single string MLM problem is defined as the problem of finding a *stochastic matrix M* satisfying the input constraint, which maximizes $M(w)$ on the input string w.

Definition 2.7 Single-String MLM Problem for C

Input: A constraint $C = \langle \Sigma, S, I, G \rangle \in \mathcal{C}$ and a string w in Σ^*.

Output: A stochastic matrix M^* satisfying G which assigns the maximum generation probability on w among all such stochastic matrices, i.e.

$$M^*(w) = max\{M(w) : M \in \mathcal{M}(G)\}$$

As usual let **P** denote the class of decision problems decidable in polynomial time and **NP** the class of decision problems acceptable in non-deterministic polynomial time. **RP** denotes the class of decision problems that are acceptable in random polynomial time (Gill, 1977): A decision problem L is said to be accepted in random polynomial time if and only if there exists a randomized algorithm A, that is, A has access to a fair coin, such that A halts in polynomial time on all inputs, and A always outputs 'no' on a negative instance and outputs 'yes' with probability at least a half on a positive instance. It is widely conjectured that **P** is strictly contained in **NP**, and also that **RP** is strictly contained in **NP**. All hardness results of this paper only hold modulo one of the above conjectures.

3. Sample complexity bounds for training PAs

Our main positive result on the training problem is the following bound on the sample complexity of the PA training problem.

Theorem 3.1. *An arbitrary class of PA constraints C is trainable with sample complexity $O((n/\epsilon)^2 t \cdot \log^3 nt/\epsilon \cdot \log 1/\delta \cdot \log^2 \log 1/\delta)$, where t is the size of the input constraint.*

Note that the above sample complexity bound is essentially linear in the size of the input constraint t, and a low-order polynomial in n, $1/\epsilon$, and $\log 1/\delta$. As an easy corollary, the following bound on the sample complexity of the training problem for the null constraints follows.

Corollary 3.1. *The class of null PA constraints is trainable with sample size: $O((n/\epsilon))^2 s^2 a \cdot \log^3 nsa/\epsilon \cdot \log 1/\delta \cdot \log^2 \log 1/\delta)$, where s is the number of states and a is the alphabet size of the null constraint to be trained.*

Outline of the proof of Theorem 3.1

Let an input constraint $C = \langle \Sigma, S, I, G \rangle \in \mathcal{C}$ be given, and let t be its size, i.e. $t = |I| + |G|$. Here I is a subset of the set S of all states, and $G \subseteq S \times S \times \Sigma$. Assume that $min\{d_{KL}(D, P) : P \in \mathcal{PA}(C)\}$ is finite with respect to the target probability distribution D, since if the minimum is infinite then by the definition of trainability any sample complexity suffices. Our objective is to show that there exists a training algorithm such that for any sufficiently large sample Ξ, its output $Q \in \mathcal{PA}(C)$ is likely to approximately minimize $d_{KL}(D, Q)$, where D is the source distribution. Recall that $d_{KL}(D, Q) = E_D(\log 1/Q(\cdot)) - H(D)$. Since the second term (the entropy of D) is independent of Q, in order to find a Q that minimizes $d_{KL}(D, Q)$, it suffices to find a Q that minimizes $E_D(\log 1/Q(\cdot))$. Thus a natural attempt would be to find a Q that minimizes $E_{\hat{D}_{\bar{z}}}(\log 1/Q(\cdot))$ and show that the empircial estimates converge to their true means *uniformly* for the class of random variables $\mathcal{F}(\mathcal{PA}(C)) = \{\log 1/P(\cdot) : P \in \mathcal{PA}(C)\}$ for moderate sample size. The difficulty here is the fact that $\mathcal{F}(\mathcal{PA}(C))$ is unbounded in the sense that $\log 1/P(x)$ diverges to infinity with $P(x)$ goes to zero. This fact prohibits the direct application of certain lemmas on the convergence of bounded random variables, such as Hoeffding's inequality.

It turns out, however, that we do not need to show uniform convergence for the entire class. We let our training algorithm output a probabilistic automaton Q for which $E_{\hat{D}_{\vec{z}}}(\log 1/Q(\cdot))$ is minimized when Q is restricted to be in a finite subclass $\mathcal{FPA}(C)_m$ of $\mathcal{PA}(C)$. (If there is more than one member in the finite class that achieves the minimum we make no further assumption about how our algorithm picks its hypothesis.) To describe $\mathcal{FPA}(C)_m$, first note that $\mathcal{PA}(C)$ is nothing but the subset of the 'parameter space' $[0, 1]^I \times [0, 1]^G$ satisfying the stochastic condition. For convenience we redefine $\mathcal{PA}(C)$ by relaxing the equalities in the stochastic condition to inequalities:

$$\forall i \in S \sum_{j \in S, z \in \Sigma} M_P(i, j, z) \leq 1 \text{ and } \sum_{j \in I} \pi_P(j) \leq 1 \tag{3.1}$$

Note that any member P of the parameter space satisfying this weaker condition can easily be converted to a probabilistic automaton P' satisfying the strict stochastic condition which assigns at least as large a probability on every word as P:

$$\forall i, j \in S, z \in \Sigma \; M_{P'}(i, j, z) = \frac{M_P(i, j, z)}{\Sigma_{k \in S, z' \in \Sigma} M_P(i, k, z')} \text{ and } \pi_{P'}(i) = \frac{\pi_P(i)}{\Sigma_{j \in I} \pi_P(j)}$$

Let us say that P' is the *stochastic correction* of P. Note that P' has less divergence than P with respect to *any* distribution[7].

$\mathcal{FPA}(C)_m$ is defined for each sample size m and is the set of stochastic corrections of all 'bounded grid points' of $\mathcal{PA}(C)$, that is, those members of $\mathcal{PA}(C)$ satisfying (3.1), in which all transition and initial probabilities are bounded from below by some decreasing function of m and are powers of $(1 - \gamma)$ where γ also is a decreasing function of m.

Now, since $\mathcal{F}(\mathcal{FPA}(C)_m) = \{\log 1/F(\cdot) : F \in \mathcal{FPA}(C)_m\}$ is a finite class of bounded random variables of moderate cardinality, we can show *fast* uniform convergence (of empirical estimates to true means) for them. We then show that for sufficiently large m, for each P in $\mathcal{PA}(C)$, there exists $F \in \mathcal{FPA}(C)_m$ which is 'close to' P everywhere in the domain:

$$\forall P \in \mathcal{PA}(C) \; \exists F \in \mathcal{FPA}(C)_m \; \forall x \in \Sigma^n \log \frac{1}{F(x)} - \log \frac{1}{P(x)} \leq \frac{2(n + 1)}{m - 1} \tag{3.2}$$

This immediately implies that for an arbitrary source distribution D over Σ^n, we have:

$$\forall P \in \mathcal{PA}(C) \; \exists F \in \mathcal{FPA}(C)_m \; d_{KL}(D, F) - d_{KL}(D, P) \leq \frac{2(n + 1)}{m - 1} \tag{3.3}$$

So if we let *Opt* be a member of $\mathcal{PA}(C)$ satisfying $d_{KL}(D, Opt) = min\{d_{KL}(D, P) : P \in \mathcal{PA}(C)\}$, then we have:

$$\exists F \in \mathcal{FPA}(C)_m \; d_{KL}(D, F) - d_{KL}(D, Opt) \leq \frac{2(n + 1)}{m - 1}$$

Then if we let F^* be a member of $\mathcal{FPA}(C)_m$ such that for all $F \in \mathcal{FPA}(C)_m$, $d_{KL}(D, F^*) \leq d_{KL}(D, F)$, then we have:

$$d_{KL}(D, F^*) - d_{KL}(D, Opt) \leq \frac{2(n+1)}{m-1} \tag{3.4}$$

Suppose that m is at least $4(n+1)/\epsilon + 1$. Thus $2(n+1)/m - 1 \leq \epsilon/2$ and

$$d_{KL}(D, F^*) - d_{KL}(D, Opt) \leq \frac{\epsilon}{2} \tag{3.5}$$

Now by the uniform convergence for the random variables $\mathcal{F}(\mathcal{FPA}(C)_m)$ as stated in Lemma 3.1, for moderately large m ($m \geq q(4/\epsilon, \log 1/\delta, n, t)$ where q is polynomially bounded and will be specified in Lemma 3.1), we have with probability at least $1 - \delta$:

$$\forall F \in \mathcal{FPA}(C)_m \left| E_{\hat{D}_{\bar{z}}}\left[\log \frac{1}{F(\cdot)}\right] - E_D\left[\log \frac{1}{F(\cdot)}\right] \right| \leq \frac{\epsilon}{4} \tag{3.6}$$

Now let $Q \in \mathcal{FPA}(C)_m$ be the output of our training algorithm, minimizing the empirical estimate of $\log 1/Q(\cdot)$ over $Q \in \mathcal{FPA}(C)_m$. Then, by (3.6), we have that with probability at least $1 - \delta$, both of the following hold:

$$E_D\left[\log \frac{1}{Q(\cdot)}\right] - E_{\hat{D}_{\bar{z}}}\left[\log \frac{1}{Q(\cdot)}\right] \leq \frac{\epsilon}{4} \text{ and} \tag{3.7}$$

$$E_{\hat{D}_{\bar{z}}}\left[\log \frac{1}{F^*(\cdot)}\right] - E_D\left[\log \frac{1}{F^*(\cdot)}\right] \leq \frac{\epsilon}{4} \tag{3.8}$$

Also by the definition of our training algorithm

$$E_{\hat{D}_{\bar{z}}}\left[\log \frac{1}{Q(\cdot)}\right] - E_{\hat{D}_{\bar{z}}}\left[\log \frac{1}{F^*(\cdot)}\right] \leq 0 \tag{3.9}$$

By summing the inequalities, (3.7), (3.8) and (3.9), we have the following with probability at least $1 - \delta$:

$$E_D\left[\log \frac{1}{Q(\cdot)}\right] - E_D\left[\log \frac{1}{F^*(\cdot)}\right] \leq \frac{\epsilon}{2} \tag{3.10}$$

This implies:

$$d_{KL}(D, Q) - d_{KL}(D, F^*) \leq \frac{\epsilon}{2} \tag{3.11}$$

Then from (3.5) and (3.11) we have with probability at least $1 - \delta$:

$$d_{KL}(D, Q) - d_{KL}(D, Opt) \leq \epsilon \tag{3.12}$$

Thus any training algorithm that minimizes the empirical divergence over $\mathcal{GPA}(C)_m$ also approximately minimizes the actual divergence over $\mathcal{PA}(C)$ within accuracy ϵ with probability $1 - \delta$, whenever $m \geq max\{4(n + 1)/\epsilon + 1, q(4/\epsilon, \log 1/\delta, n, t)\}$.

Proof of Theorem 3.1

We need to show that (3.2) holds and that whenever $m \geq q(4/\epsilon, \log 1/\delta, n, t)$ the inequality (3.6) holds with probability at least $1 - \delta$, for some q for which $q(4/\epsilon, \log 1/\delta, n, t)$ is at least $4(n + 1)/\epsilon + 1$ and which fulfills the order bound promised in the statement of the theorem.

We begin by defining $\mathcal{GPA}(C)_m$. We define a finite subset (grid points), denoted by $\mathcal{G}(\gamma, \theta)$, of $\mathcal{PA}(C)$ as the set of all members P in $\mathcal{PA}(C)$ satisfying the relaxed stochastic condition (3.1), such that each of $\pi_P(i)$ and $M_P(i, j, z)$ is some power of $1 - \gamma$, and is at least θ. Let $\mathcal{G}'(\gamma, \theta)$ be the set of stochastic corrections of the members of $\mathcal{G}(\gamma, \theta)$. We then define $\mathcal{GPA}(C)_m$ as $\mathcal{G}'(1/m, 1/2tm)$.

We next verify (3.2) in two steps (inequalities (3.13) and (3.14) below). We define $\mathcal{BPA}(C)_m$, a subclass of $\mathcal{PA}(C)$ in which all transition and initial probabilities are bounded from below by[8] $1/tm$.

$$\mathcal{BPA}(C)_m = \{B \in \mathcal{PA}(C) : \forall i, j \in S, z \in \Sigma\ M_B(i, j, z) \geq \frac{1}{tm} \text{ and } \pi_B(i) \geq \frac{1}{tm}\}$$

We then show that for arbitrary P in $\mathcal{PA}(C)$, there exists $B \in \mathcal{BPA}(C)_m$ which is close to P in the following sense:

$$\forall P \in \mathcal{PA}(C)\ \exists B \in \mathcal{BPA}(C)_m\ \forall x \in \Sigma^n \log \frac{1}{B(x)} - \log \frac{1}{P(x)} \leq \frac{n + 1}{m - 1} \tag{3.13}$$

In addition, we show that for each member of $\mathcal{BPA}(C)_m$ there is one in $\mathcal{GPA}(C)_m$ that is close to it:

$$\forall B \in \mathcal{BPA}(C)\ \exists F \in \mathcal{GPA}(C)_m\ \forall x \in \Sigma^n \log \frac{1}{F(x)} - \log \frac{1}{B(x)} \leq \frac{n + 1}{m - 1} \tag{3.14}$$

(3.2) clearly follows from (3.13) and (3.14). To verify the first inequality (3.13), let an arbitrary $P \in \mathcal{PA}(C)$ be given. We obtain P_m from P by shifting each $M_P(i, j, z)$ towards the '*uniform* stochastic matrix' to obtain P_m—the matrix in which each transition in G out of any state i receives the same probability $1/t(i)$, where we let $t(i)$ denote the number of transitions out of state i. Formally:

$$\forall i, j \in S\ \forall z \in \Sigma\ M_{P_m}(i, j, z) = \frac{(m - 1) \cdot M_P(i, j, z) + (1/t(i))}{m} \tag{3.15}$$

Similarly we shift each $\pi_P(i)$ towards the uniform distribution over I:

$$\forall i \in S\ \pi_{P_m}(i) = \frac{(m - 1) \cdot \pi_P(i) + (1/|I|)}{m} \tag{3.16}$$

Note that for all i, j, z, $M_{P_m}(i, j, z) \geq 1/mt(i) \geq 1/|G|m \geq 1/tm$ *and* $\pi_{P_m}(i) \geq 1/|I|m \geq 1/tm$, and hence $P_m \in \mathcal{BPA}(C)_m$. Now since each of the initial and transition probabilities in P_m is at least as much as $(m - 1)/m$ times the corresponding probability in P, the probability assigned by P_m on any path (and hence on any string) is at least $((m - 1)/m)^{n+1}$ times that assigned by P. Hence we have for any string $x \in \Sigma^n$:

$$\frac{P(x)}{P_m(x)} \leq \left(1 + \frac{1}{m - 1}\right)^{n+1} \leq e^{n+1/m-1}$$

Hence, for every $x \in \Sigma^n$, we have

$$\log \frac{1}{P_m(x)} - \log \frac{1}{P(x)} \leq \frac{n + 1}{m - 1}$$

This completes the proof of (3.13). The second inequality (3.14) is straightforward to verify. Let B be an arbitrary member of $\mathcal{BPA}(C)_m$. We round off each initial probability and each transition probability in B to a power of $(1 - 1/m)$. Denote the obtained PA by R and its stochastic correction by F. Each probability in F is at least $(1 - 1/m)$ times the corresonding probability in B and this leads to two consequences. First, since $B \in \mathcal{BPA}(C)_m$ each initial and transition probability in F is at least $1/tm(1 - 1/m) \geq 1/2tm$, since the final choice of m will be larger than 2. This implies that the 'nearest grid point' R next to B lies in $\mathcal{G}(1/m, 1/2tm)$ and the stochastic correction F lies in $\mathcal{FPA}(C)_m = \mathcal{G}'(1/m, 1/2tm)$. Second, each probability in B is at most $(1 + 1/(m - 1))$ times the corresponding probability in F. Therefore, for any $x \in \Sigma^n$ we must have:

$$\frac{B(x)}{F(x)} \leq \left(1 + \frac{1}{m - 1}\right)^{n+1} \leq e^{n+1/m-1}$$

and hence:

$$\log \frac{1}{F(x)} - \log \frac{1}{B(x)} \leq \frac{n + 1}{m - 1}$$

This proves (3.14) and completes the proof of (3.2).

The next lemma shows that whenever $m \geq q(4/\epsilon, \log 1/\delta, n, t)$ the inequality (3.6) holds with probability at least $1 - \delta$.

Lemma 3.1. *Let* $\mathcal{F}(\mathcal{FPA}(C)_m) = \{\log 1/F(\cdot) : F \in \mathcal{FPA}(C)_m\}$. *Let* D *be an arbitrary distribution over* Σ^n *for some* $n \in N$, *and* D^m *denote the product distribution induced by* D *over* $(\Sigma^n)^m$. *If* $\Xi = \langle w_1, \ldots, w_m \rangle \in (\Sigma^n)^m$ *then we let* $E_{\hat{D}_\Xi}(f)$ *denote the empirical estimate of the random variable* f *by the sample* Ξ, *i.e.* $E_{\hat{D}_\Xi}(f) = \Sigma_{i=1}^m f(w_i)/m$, *and let* $E_D(f)$ *denote the expectation (according to* D*) of* f. *Then we have, for all* $\epsilon \leq 1$ *and* $\delta > 0$, *for all* $n, t \in N$, *whenever* $m \geq q(1/\epsilon, \log 1/\delta, n, t)$,

$$D^m \{\Xi \in (\Sigma^n)^m : \exists f \in \mathcal{F}(\mathcal{FPA}(C)_m) \text{ such that } |E_{\hat{D}_\Xi}(f) - E_D(f)| > \epsilon\} < \delta,$$

where $q(1/\epsilon, \log 1/\delta, n, t)$ *is defined as*

$$max\left\{\frac{32(n+1)^2t}{\epsilon^2}\ln^3\frac{64t(n+1)^2t}{\epsilon^2}, \frac{8(n+1)^2\ln 1/\delta}{\epsilon^2}\log^2\frac{8t(n+1)^2\ln 1/\delta}{\epsilon^2}\right\}$$

Proof of Lemma 3.1. We use the following lemma which follows from Hoeffding's inequality. (See for example Pollard (1984).)

Lemma 3.2 (Hoeffding). *Let* $\mathcal{F}$ *be a finite class of bounded random variables on a set* X, *that is for each* $f \in \mathcal{F}$, $f : X \to [0, M]$ *for some real* $M \in R$. *Let* D *be an arbitrary distribution over* X. *Then we have:*

$$\text{If } m \geq \frac{M^2}{\epsilon^2}\left(\ln|\mathcal{F}| + \ln\frac{1}{\delta}\right)$$

$$\text{then } D^m \{\Xi \in X^m : \exists f \in \mathcal{F} \mid E_{\hat{D}_\Xi}(f) - E_D(f)| > \epsilon\} < \delta$$

To apply Lemma 3.2, we compute an upper bound on the random variables in $\mathcal{F}(\mathcal{FPA}(C)_m)$, and the cardinality of $\mathcal{FPA}(C)_m$. Since by our assumption at the beginning of the proof $min\{d_{KL}(D, P) : P \in \mathcal{PA}(C)\}$ is finite, for any string x in Σ^n assigned a positive probability by the target distribution D, there is a path for x in the input constraint C. Now any path in C is produced with probability at least $(1/2tm)^{n+1}$, since any state is chosen as the inital state with probability at least $1/2tm$ and each transition in the path has probability at least $1/2tm$. Hence $\log 1/F(x)$ is bounded from above by $(n+1)\log 2tm$. Since $\mathcal{FPA}(C)_m$ equals $\mathcal{G}'(1/m, 1/2tm)$, and the cardinality of $\mathcal{G}'(\gamma, \theta)$ is at most $(1/\gamma \ln 1/\theta)^t$, the cardinality of $\mathcal{FPA}(C)_m$ is at most $(m \ln 2tm)^t$. Plugging in $M = (n+1)\log 2tm$ and $|\mathcal{FPA}(C)_m| \leq (m \ln 2tm)^t$ into the inequality in Lemma 3.2 gives the following inequality:

$$m \geq \frac{(n+1)^2\log^2 2tm}{\epsilon^2}\left(t\ln(m\ln 2tm) + \ln\frac{1}{\delta}\right) \tag{3.17}$$

To show that if m is at least the bound in Lemma 3.1 then (3.17) holds, it suffices to show that the following two inequalities hold under the same condition.

$$\frac{m}{2} \geq \frac{(n+1)^2 \log^2 2tm}{\epsilon^2} t \ln (m \ln 2tm) \tag{3.18}$$

$$\frac{m}{2} \geq \frac{(n+1)^2 \log^2 2tm}{\epsilon^2} \ln \frac{1}{\delta} \tag{3.19}$$

To get a simple argument for (3.18) we first show that

$$\log^2 2tm \ln (m \ln 2tm) \leq \frac{\ln^2 2tm}{\ln^2 2} \ln (tm \ln 2tm) \leq 2 \ln^3 2tm \tag{3.20}$$

For $tm \geq 1$ the latter inequality is equivalent to $\ln \ln 2tm \leq (2 \ln^2 2 - 1) \ln 2tm + \ln 2$ which holds for $tm = 1$ and only improves for larger m. We now return to the proof of (3.18): Since (3.20) holds it suffices to show

$$m \geq \alpha \ln^3 \beta m, \text{ for } \alpha = \frac{4(n+1)^2 t}{\epsilon^2} \text{ and } \beta = 2t. \tag{3.21}$$

If $m \geq \alpha \ln^3 \beta m$ holds for some choice of m then it also does for all larger m. We set m to $8\alpha \ln^3 8\alpha\beta$ which is the first bound in the maximization clause of $q(1/\epsilon, \log 1/\delta, n, t)$. Then $m \geq \alpha \ln^3 \beta m$ is equivalent to $8\alpha\beta \geq \ln^3 8\alpha\beta$. Since $n, t \geq 1$, $\alpha\beta = 4(n+1)^2 t/\epsilon^2 \cdot (2t)$ is always at least 32. Now observe that the last inequality holds for $\alpha\beta = 32$ because $(\ln 256)^3 \approx 170.51 < 256$, and thus for all larger $\alpha\beta$. We conclude that (3.21) and hence (3.18) holds.

The proof of (3.19) is simpler. We need to show

$$m \geq \alpha \log^2 \beta m, \text{ for } \alpha = \frac{2(n+1)^2}{\epsilon^2} \ln \frac{1}{\delta} \text{ and } \beta = 2t.$$

One can show that if $m \geq 4\alpha \log^2 2\alpha\beta$ (which is the second bound in the maximization clause of $q(1/\epsilon, \log 1/\delta, n, t)$) then $m \geq \alpha \log^2 \beta m$.

To complete the proof of Theorem 3.1 observe that $q(4/\epsilon, \log 1/\delta, n, t)$ of Lemma 3.1 is at least $4(n+1)/\epsilon + 1$ and fulfills the order bound promised in the statement of the theorem. □

As an immediate corollary to the proof of Theorem 3.1, we have the following positive result for training PAs of a fixed number of parameters.

Corollary 3.2. *Any finite class of PA constraints is polynomially trainable.*

Proof of Corollary 3.2

Since the sample size m required in Theorem 3.1 is polynomial in $1/\epsilon$, $1/\delta$, n and t, this immediately follows from the observation that $| \mathcal{FPA}(C)_m | \leq (m \log 2tm)^t$ is polynomial in m when t is bounded from above by a constant. □

4. The equivalence between the training and maximum likelihood model problems

The sample complexity of Theorem 3.1 can be used to establish the equivalence between the efficient trainability of a class of cosntraints and the efficient approximability of the sample MLM problem for the same class. We first define what it means for a randomized algorithm to *approximate* the sample MLM problem within a given factor.

Definition 4.1 (Approximate Sample MLM problem). *A randomized algorithm A is said to approximate the sample MLM problem for a class of constraints C within factor K, possibly a function of various parameters of the problem, in random $T(\ldots)$ time, if given a constraint $C \in \mathcal{C}$ and an input sample $\Xi = \langle w_1, \ldots, w_m \rangle$ of Σ^n for some $n > 0$ and some finite alphabet Σ, A terminates in $T(\ldots)$ many steps and outputs a PA $Q \in \mathcal{PA}(C)$, which with probability at least a half satisfies:*

$$\frac{\Pi_{i=1}^m OPT(w_i)}{\Pi_{i=1}^m Q(w_i)} \leq K$$

where OPT is a member of $\mathcal{PA}(C)$ which maximizes the likelihood of Ξ, i.e.

$$\prod_{i=1}^m OPT(w_i) = max \left\{ \prod_{i=1}^m P(w_i) : P \in \mathcal{PA}(C) \right\}$$

By convention, we let $0/0 = 1$

As before, *OPT* is guaranteed to exist because of the compactness of $\mathcal{PA}(C)$ and the continuity of the likelihood function on a finite sample.

Theorem 4.1. *For an arbitrary class of PA constraints C, the following four statements are equivalent. Below, we let t denote the size of the input constraint $C \in \mathcal{C}$ to be trained, m the sample size, and n the length of each example.*

1. *There exists a training algorithm for C with sample complexity polynomial in $1/\epsilon$, $1/\delta$, t and n, running in time polynomial in the total sample length.*
2. *There exists a training algorithm for C with sample complexity polynomial in $1/\epsilon$, $\log 1/\delta$, t and n, running in time polynomial in the total sample length.*
3. *The sample MLM problem for C is approximable within a factor of $1 + \epsilon$ in random time polynomial in $1/\epsilon$, t, n and m.*

4. *The sample MLM problem for C is approximable within a factor of $2^{p(n,t)m^{\alpha}}$ in random time polynomial in t, n and m, for some polynomial p and $\alpha < 1$.*

Proof of Theorem 4.1

(1 $\rightarrow$ 3) The idea of the proof is as follows. We use the hypothetical training algorithm A for a class of constraints C to construct a randomized approximation algorithm B for the sample MLM problem for the same class.

In order to do this, we take advantage of the *robustness* of the training algorithm. In particular, the algorithm must meet its performance guarantee when it is fed examples generated by the *empirical distribution* observed in the input sample Ξ for the sample MLM problem. We then appeal to the classical equivalence between minimizing the divergence with respect to the empirical distribution and maximizing the likelihood of a given sample (see Lemma 2.1).

Let $\epsilon > 0$, a finite sample $\Xi = \langle w_1, \ldots, w_m \rangle$ with $|w_i| = n$ for each w_i, and a constraint $C \in \mathcal{C}$ be given. We use the hypothetical training algorithm A on the empirical distribution $\hat{D}_{\Xi}$ of Ξ, as defined in Definition 2.1. We can assume without loss of generality that $min\{d_{KL}(\hat{D}_{\Xi}, P) : P \in \mathcal{PA}(c)\}$ is finite, since otherwise $max\{\Pi_{i=1}^{m} P(w_i) : P \in \mathcal{PA}(C)\} = 0$ and the MLM problem is trivial. B then computes a sample size m' for A large enough for accuracy ϵ/m and confidence $1 - 1/4$. For example, $m' = \lceil q(m/\epsilon, 4, n, t) \rceil$ suffices where $q(\ldots)$ is an upper bound on the sample complexity of the hypothetical training algorithm A, which by assumption is polynomially bounded. Now, B gives A a sample Ξ' of size m' obtained by sampling from Ξ according to $\hat{D}_{\Xi}$, the empirical distribution of Ξ, or the uniform distribution over the m elements of the sequence Ξ. Here note that while it is not always possible to simulate the empirical distribution of a finite sample with a fair coin, there exists an algorithm which generates a new sample of an arbitrary size acording to the empirical distribution of the original sample with high probability. We make this precise below.

Lemma 4.1. *There exists a randomized algorithm, U, which, given a finite sample $\Xi = \{w_1, \ldots, w_m\}$ of size m, an integer m', and a confidence parameter $\nu > 0$, always terminates in time polynomial in m, m' and $1/\nu$ and outputs a sample Ξ' of size m' which with probability at least $1 - \nu$ is drawn according to the empirical distribution $\hat{D}_{\Xi}$ of Ξ.*

Proof of Lemma 4.1. The algorithm U first calculates $i = \lceil \log m' \rceil$ and iterates the following: U flips a fair coin i times to obtain a bit string x of length i. If $x < m$, then U appends w_{x+1} to the end of the sequence Ξ', and does nothing otherwise. It is clear that each example of Ξ' is drawn independently at random according to $\hat{D}_{\Xi}$. Notice that at each iternation the length of Ξ' increases by one with probability at least a half. It is easy to see, by applying Chernoff's bound (c.f. Valiant (1984)) that in $p(m', 1/\nu)$ many iterations, the length of Ξ' becomes m' with probability at least $1 - \nu$, where p is some polynomial. If this fails to occur, i.e. the length of Ξ' is shorter than m' after $p(m', 1/\nu)$ many iterations, U pads Ξ' with arbitrary examples to make its length m'.

End of proof of Lemma 4.1

We run U on Ξ, m' and 1/3 to obtain a sample Ξ', and let Q be the output obtained by running A on Ξ'. Thus, A is run on a sample of size m' generated with respect to $\hat{D}_\Xi$ with probability at least 2/3. Furthermore, when this is the case, by the performance guarantee on A, the divergence of Q with respect to $\hat{D}_\Xi$ is ϵ/m close to the divergence of the best PA satisfying C, with probability at least 3/4. Hence, with probability at least $3/4 \cdot 2/3 = 1/2$, the following holds:

$$d_{KL}(\hat{D}_\Xi, P) - d_{KL}(\hat{D}_\Xi, OPT) = \sum_{x \in \Sigma^n} \hat{D}_\Xi(x) \log \frac{OPT(x)}{P(x)} \leq \frac{\epsilon}{m} \tag{4.1}$$

where OPT is a PA in $\mathcal{PA}(C)$ satisfying:

$$d_{KL}(\hat{D}_\Xi, OPT) = \min \{d_{KL}(\hat{D}_\Xi, P) : P \in \mathcal{PA}(C)\}$$

Substituting $\hat{D}_\Xi(x) = \#(x, \Xi)/m$ in (4.1) above,

$$d_{KL}(\hat{D}_\Xi, Q) - d_{KL}(\hat{D}_\Xi, OPT) = \frac{1}{m} \sum_{i=1}^{m} \log \frac{OPT(w_i)}{Q(w_i)} \leq \frac{\epsilon}{m}$$

We thus obtain $\Sigma_{i=1}^m \log OPT(w_i)/Q(w_i) \leq \epsilon$. So for $\epsilon \in (0, 1]$, we have

$$\prod_{i=1}^{m} \frac{OPT(w_i)}{Q(w_i)} \leq 2^\epsilon \leq 1 + \epsilon$$

Since OPT is a probabilistic automaton in $\mathcal{PA}(C)$ minimizing the divergence with respect to $\hat{D}_\Xi$ on Ξ, by the 'classical equivalence' given in Lemma 2.1, it is also an optimal solution in $\mathcal{PA}(C)$ of the sample MLM problem on Ξ. Thus we have obtained a $1 + \epsilon$ approximation to the sample MLM problem with probability at least a half. The running time of B is clearly bounded by a polynomial in $1/\epsilon$, n, t, m, since m' is polynomial in these parameters, the time spent on generating the sample Ξ' using the algorithm U is polynomial in n, m and m', and the running time of A is polynomial in the total length of the sample Ξ', i.e. polynomial in m' and n.

(3 → 2)

We will use the hypothetical approximation algorithm for the MLM problem for a class of constraints C to construct B that trains C. Let D be the target distribution over Σ^n and let a constraint $C \in \mathcal{C}$ be given as input. Assume without loss of generality that $min\{d_{KL}(D, P) : P \in \mathcal{PA}(C)\}$ is finite, since otherwise the training problem is trivial. The finiteness of $min\{d_{KL}(D, P) : P \in \mathcal{PA}(C)\}$ guarantees that for each $x \in \Sigma^n$ assigned a positive probability by D, there is a path in C labeled with x.

Hence it follows that for an arbitrary sample Ξ generated by D, $max\{\Pi_{i=1}^m P(w_i) : P \in \mathcal{PA}(C)\}$ is positive. First we show that any approximation algorithm (in the sense of

Theorem 4.1, part 3) can be modified so as to output a PA Q_B in the finite class $\mathcal{FPA}(C)_m$, defined for C as in the proof of Theorem 3.1, which with probability at least $1 - \delta/2$ satisfies:

$$E_{\hat{D}_{\Xi}}\left[\log \frac{1}{Q_B(\cdot)}\right] - E_{\hat{D}_{\Xi}}\left[\log \frac{1}{OPT(\cdot)}\right] \leq \frac{3(n+1)}{m-1} \tag{4.2}$$

where $OPT \in \mathcal{PA}(C)$ is an optimal solution for the sample MLM problem with input sample Ξ satisfying the input constraint C. Second, we show that in fact such an algorithm leads to a training algorithm for PAs with a slightly larger sample complexity than the training algorithm exhibited in the proof of Theorem 3.1.

It is easy to verify the first of these two claims. First note that any algorithm which approximates the sample MLM problem for C within factor $1 + \epsilon$ with probability at least a half in time polynomial in $1/\epsilon$, n, t and m can be "boosted" to one which achieves the same approximation factor with probability at least $1 - \delta/2$ in time polynomial in $\log 2/\delta$, $1/\epsilon$, t, n and m. This can be done by iteratively running the former algorithm $\lceil \log 2/\delta \rceil$ times and then selecting, from among its outputs, one that assigns the maximum likelihood on the input sample. (Note that using dynamic programming it is easy to compute for a given sample and PA the likelihood of that sample in time polynomial in the total length of the sample and t.) Now set $\epsilon = n$ and run this boosted algorithm on the input sample S and obtain a PA P, which with probability at least $1 - \delta/2$ approximates the sample MLM problem for C within a factor of $1 + n$:

$$\prod_{i=1}^{m} \frac{OPT(w_i)}{P(w_i)} \leq 1 + n \leq e^{1+n}$$

Hence we have:

$$E_{\hat{D}_{\Xi}}\left[\log \frac{1}{P(\cdot)}\right] - E_{\hat{D}_{\Xi}}\left[\log \frac{1}{OPT(\cdot)}\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m} \log \frac{OPT(w_i)}{P(w_i)} \leq \frac{n+1}{m} < \frac{n+1}{m-1} \tag{4.3}$$

Now, using the trick of shifting, rounding-off and stochastic correction as done in the proof of Theorem 3.1, we obtain from P a member Q_B of $\mathcal{FPA}(C)_m$, such that

$$\forall x \in \Sigma^n \log \frac{1}{Q_B(x)} - \log \frac{1}{P(x)} \leq \frac{2(n+1)}{m-1} \tag{4.4}$$

From (4.3) and (4.4), (4.2) follows.

Next we can show that the boosted approximation algorithm for the sample MLM problem indeed trains C. This can be shown similarly to the way the algorithm exhibited in the proof of Theorem 3.1 was proved to successfully train C. Let us first recall the notation of that proof: We denoted by Q the output of the training algorithm which, for the input sample Ξ, minimized $E_{\hat{D}_\Xi}(\log 1/F(\cdot))$ over $F \in \mathcal{FPA}(C)_m$. The PA Opt was a member of $\mathcal{PA}$ with the minimum divergence with respect to the target distribution. F^* denoted a member of $\mathcal{FPA}(C)_m$ that is closest to Opt. In order to prove that $E_D(\log 1/Q(\cdot))$ is close to $E_D(\log 1/Opt(\cdot))$ with high probability (inequality (3.12)), we showed in the proof of Theorem 3.1 that (inequality (3.4)):

$$E_D\left[\log \frac{1}{F^*(\cdot)}\right] - E_D\left[\log \frac{1}{Opt(\cdot)}\right] \leq \frac{2(n+1)}{m-1} \tag{4.5}$$

and whenever $m \geq q(4/\epsilon, 1/\delta, n, t)$ with probability at least $1 - \delta$ (inequality (3.10)),

$$E_D\left[\log \frac{1}{Q(\cdot)}\right] - E_D\left[\log \frac{1}{F^*(\cdot)}\right] \leq \frac{\epsilon}{2} \tag{4.6}$$

Since (4.5) always holds, we only need to show an analogue of (4.6) for any algorithm satisfying (4.2). Recall that in the earlier proof, in order to show (3.10), we applied triangle inequality to the inequalities (3.7), (3.8), and (3.9). (3.7) and (3.8) followed from the uniform convergence for $\mathcal{FPA}(C)_m$, and (3.9) followed from the fact that the training algorithm minimized $E_{\hat{D}_\Xi}(\log 1/Q(\cdot))$ within $\mathcal{FPA}(C)_m$.

In this proof, the output Q_B of our training algorithm *approximately* minimizes $E_{\hat{D}_\Xi}(\log 1/Q(\cdot))$ within $\mathcal{FPA}(C)_m$. More precisely, by the optimality of OPT, it follows from (4.2) that the following holds with probability at least $1 - \delta/2$:

$$E_{\hat{D}_\Xi}\left[\log \frac{1}{Q_B(\cdot)}\right] - E_{\hat{D}_\Xi}\left[\log \frac{1}{F^*(\cdot)}\right] \leq \frac{3(n+1)}{m-1} \tag{4.7}$$

As before, for $m \geq q(4/\epsilon, \log 2/\delta, n, t)$ both of the following hold with probability at least $1 - \delta/2$:

$$E_D\left[\log \frac{1}{Q_B(\cdot)}\right] - E_{\hat{D}_\Xi}\left[\log \frac{1}{Q_B(\cdot)}\right] \leq \frac{\epsilon}{4} \tag{4.8}$$

$$E_{\hat{D}_\Xi}\left[\log \frac{1}{F^*(\cdot)}\right] - E_D\left[\log \frac{1}{F^*(\cdot)}\right] \leq \frac{\epsilon}{4} \tag{4.9}$$

By summing up the inequalities (4.7), (4.8), and (4.9), we obtain that the following holds with probability at least $1 - \delta$ whenever $m \geq q(4/\epsilon, \log 2/\delta, n, t)$:

$$E_D\left[\log\frac{1}{Q_B(\cdot)}\right] - E_D\left[\log\frac{1}{F^*(\cdot)}\right] \le \frac{\epsilon}{2} + \frac{3(n+1)}{m-1} \tag{4.10}$$

Hence from (4.5) and (4.10), it follows that whenever $m \ge q(4/\epsilon, \log 2/\delta, n, t)$, with probability at least $1 - \delta$,

$$E_D\left[\log\frac{1}{Q_B(\cdot)}\right] - E_D\left[\log\frac{1}{Opt(\cdot)}\right] \le \frac{\epsilon}{2} + \frac{5(n+1)}{m-1} \tag{4.11}$$

Hence, whenever $m \ge max\{q(\epsilon/4, \log 2/\delta, n, t), 10(n+1)/\epsilon + 1\}$, with probability at least $1 - \delta$,

$$E_D\left[\log\frac{1}{Q_B(\cdot)}\right] - E_D\left[\log\frac{1}{Opt(\cdot)}\right] \le \epsilon$$

We have thus shown that the above algorithm trains C, which was an arbitrary constraint in G, with sample complexity polynomial in $1/\epsilon$, $\log 1/\delta$, n, t and running time polynomial in the same parameters and m. It is easy to convert such an algorithm to one whose running time is polynomial in the total sample length and which has sample complexity still polynomial in $1/\epsilon$, $\log 1/\delta$, n, t.

(2 → 1)

This is obvious from the definitions.

Thus far, we have shown that the first three statements of the thorem are equivalent. Now we proceed to show that *3* is equivalent to *4*.

(4 → 3) Suppose that algorithm A, running in random time polynomial in $1/\epsilon$, n, t, m, approximates the sample MLM problem for G within factor $2^{p(n,t)m^\alpha}$ for some polynomial p. Let a constraint $C \in G$ and a sample $\Xi = \langle w_1, \ldots, w_m\rangle$ be given. Again, assume without loss of generality that $\Pi_{i=1}^m OPT_\Xi(w_i) = max\{\Pi_{i=1}^m P(w_i) : P \in \mathcal{PA}(C)\}$ is positive. We then repeat the sample $r = \lceil(2/\epsilon(\ln 2)p(n,t)m^\alpha)^{1/1-\alpha}\rceil$ times to obtain a new sample $\Xi' = \langle v_1, \ldots, v_{rm}\rangle$ of length rm, and feed this into A to obtain a hypothesis Q. Then, by definition, Q must satisfy:

$$\frac{\Pi_{i=1}^{rm} OPT_{\Xi'}(t_i)}{\Pi_{i=1}^{rm} Q(t_i)} \le 2^{p(n,t)(rm)^\alpha}$$

where we used $OPT_{\Xi'}$ to denote the PA in $\mathcal{PA}(C)$ that assigns the maximum probability on the sample Ξ'. Since each example of S is repeated exactly r times in Ξ',

$$\frac{\Pi_{i=1}^{rm} OPT_{\Xi'}(w_i)}{\Pi_{i=1}^{rm} Q(w_i)} \le (2^{p(n,t)(mr)^\alpha})^{1/r}$$

$$\le 2^{p(n,t)m^\alpha r^{\alpha-1}} \tag{4.12}$$

By substituting r into the exponent in (4.12) we obtain:

$$\log \frac{\Pi_{i=1}^{rm} OPT_{\Xi'}(w_i)}{\Pi_{i=1}^{rm} Q(w_i)} \leq p(n, t)m^{\alpha} \left\lceil \left(\frac{2}{\epsilon} (\ln 2) p(n, t)m^{\alpha}\right)^{1/1-\alpha} \right\rceil^{\alpha-1}$$

$$\leq p(n, t)m^{\alpha} \left(\left(\frac{2}{\epsilon} (\ln 2) p(n, t)m^{\alpha}\right)^{1/1-\alpha}\right)^{\alpha-1}, \text{ since } \alpha - 1 < 0$$

$$= p(n, t)m^{\alpha} \left(\frac{2}{\epsilon} (\ln 2) p(n, t)m^{\alpha}\right)^{-1}$$

$$= \frac{\epsilon}{2} \log e$$

Hence, we have:

$$\frac{\Pi_{i=1}^{m} OPT_{\Xi'}(w_i)}{\Pi_{i=1}^{m} Q(w_i)} \leq 2^{\epsilon/2 \log e}$$

$$= e^{\epsilon/2}$$

$$\leq 1 + \epsilon, \text{ for } \epsilon \in (0, 1]. \tag{4.13}$$

We next show that OPT_{Ξ} and $OPT_{\Xi'}$ assign the same likelihood on Ξ:

$$\prod_{i=1}^{m} OPT_{\Xi}(w_i) = max\{\prod_{i=1}^{m} P(w_i) : P \in \mathcal{PA}(C)\}$$

$$= (max\{\prod_{i=1}^{mr} P(v_i) : P \in \mathcal{PA}(C)\})^{1/r}$$

$$= (\prod_{i=1}^{mr} OPT_{\Xi'}(v_i))^{1/r}$$

$$= \prod_{i=1}^{m} OPT_{\Xi'}(w_i) \tag{4.14}$$

By plugging (4.14) in (4.13), we finally obtain:

$$\frac{\Pi_{i=1}^{m} OPT_{\Xi}(w_i)}{\Pi_{i=1}^{m} Q(w_i)} \leq 1 + \epsilon$$

(3 → 4) This is obvious from the definitions. □

From Theorem 4.1 and Corollary 3.2, the following positive result on the approximation problem for the sample MLM problem follows at once.

Corollary 4.1. *The sample MLM problem for any finite class of PA constraints is approximable within a factor of* $1 + \epsilon$ *for any* $\epsilon > 0$, *in time polynomial in* $1/\epsilon$ *and the total length of the sample.*

Note that this does not give rise to a practical training algorithm, since the running time of the algorithm we exhibit is exponential in the number of probability parameters specified by the input constraint, which may grow quite large in practical applications.

We mention a simple class of constraints which can be trained efficiently. A constraint $C = \langle \Sigma, S, I, G \rangle$ is said to be *deterministic*, if I contains exactly one start state i_0 and if for a given state i and a given letter a, there is at most one transition from i labeled with a, in G. It is well known in the literature that the MLM problem for the class of deterministic constraints is solvable in polynomial time. We repeat the proof of this fact briefly, as it makes use of a lemma that will be used again in the hardness proof in Section 5. Let a deterministic constraint $C = \langle \Sigma, S, \{i_0\}, G \rangle$ and a sample $\Xi = \langle w_1, \ldots, w_m \rangle$ be given. For each $w_k \in \Xi$ there is at most one path in C labeled with w_k starting at i_0. Denote by Θ the set of all such paths. If there is a word in Ξ for which there is no path in Θ, then the maximum probability assignable on S is zero and the problem is trivial. Let $\#(j, z \mid i)$ denote the total number of times the transition to state j outputting letter z was taken from state i in all paths of Θ and $\#(i)$ denote the number of times a transition was taken from state i in all paths of Θ. Clearly $\Sigma_{j \in S, z \in \Sigma} \#(j, z \mid i) = \#(i)$. Now note that for any probabilistic automaton P, we can calculate $P(\Xi) = \Pi_{i=1}^{m} P(w_i)$ as follows.

$$P(\Xi) = \prod_{i,j \in S, z \in \Sigma} M_P(i, j, z)^{\#(j,z|i)}$$

We wish to maximize the above expression, subject to the constraint:

$$\forall i \in S \quad \sum_{j \in S, z \in \Sigma} M_P(i, j, z) = 1$$

We can show that this is maximized when we define P as follows, using the following well-known technical lemma, which we state without proof. (See for example Levinson, Rabiner and Sondhi (1983).)

$$\forall i, j \in S \; \forall z \in \Sigma \; M_P(i, j, z) = \frac{\#(j, z \mid i)}{\#(i)}$$

Lemma 4.2. *A function of the form* $f(x_1, \ldots, x_n) = x_1^{a_1} \cdot x_2^{a_2} \cdot \ldots \cdot x_n^{a_n}$ *subject to the constraint* $x_1 + x_2 + \ldots + x_n = c$ *attains its maximum when* $x_i = (a_i / \Sigma_{i=1}^n a_i) \cdot c$.

From the above observation and Theorem 4.1, it follows that the class of deterministic constraints is polynomially trainable. The sample complexity obtained in this way can probably be improved significantly for this restricted case.

Corollary 4.2. *The class of deterministic PA constraints is polynomially trainable.*

5. Computational complexity of training probabilistic automata

The hardness of training the 2-state null PA constraints is shown via a strong non-approximability result for the *single string* MLM problem for the same class. We emphasize again that the training problem for the class of s-state null constraints is the natural problem of finding a near optimal probabilistic automaton of a *given number of states.* It should be noted that showing that the class of null constraints is hard to train is much more significant and also more difficult than constructing an artificial class of constraints that is hard to train.

Theorem 5.1. *For any* $\alpha > 0$, *the single string MLM problem for the class of 2-state null PA constraints is not approximable within* $2^{|w|^{1-\alpha}}$ *in time polynomial in the alphabet size a and* $| w |$, *where w is the input word, unless* **P** = **NP**.

As we noted in Section 1, Theorem 5.1 is a strong non-approximability result, since guaranteed approximation ratio of $2^{|w|}$ for the MLM problem for 2-state PAs is trivially achievable.

Theorem 5.2. *For arbitrary* $s \in N$, *the single string MLM problem for the s-state null constraints is approximable within* $s^{|w|}$ *in time polynomial in the alphabet size a and* $| w |$, *where w is the input word.*

The proof of Theorem 5.2 is simple but uses one of the technical lemmas to be stated and used in the proof of Theorem 5.1, so we defer the proof of Theorem 5.2 till after the proof of Theorem 5.1.

Corollary 5.1. *The class of 2-state null PA constraints is not polynomially trainable, unless* **RP** = **NP**.

Proof of Corollary 5.1 given Theorem 5.1

Suppose that a training algorithm A trains 2-state PA constraints in time polynomial in $1/\epsilon$, $1/\delta$, n and a. Then, by Theorem 4.1, it follows that there exists an algorithm B which approximates the MLM problem within factor $1 + \epsilon$ in random time polynomial in n, m, and a. We can then use B to solve 3-SAT in random polynomial time, using the reduction of Theorem 5.1. Thus 3-SAT, which is NP-complete, would be shown to be in **RP**, which would imply that **RP** = **NP**. □

The proof of Theorem 5.1 is lengthy, so we will begin by giving a very high level description of the proof. We will then give a proof sketch, introducing some key definitions and then give the formal proof. We reduce 3-SAT, the satisfiability problem for 3-CNF formulas, to the approximation problem for the single string MLM problem for the 2-state null constraints with a guaranteed approximation ratio of $2^{|w|^{1-\alpha}}$ for any fixed $\alpha > 0$. For an arbitrary $\alpha > 0$, we exhibit a polynomial time reduction which maps any CNF formula F to a string w such that the maximum probability assignable by a 2-state stochastic matrix on the string w is at least $C(F)$ if the formula is satisfiable, and less than $1/2^{|w|^{1-\alpha}} C(F)$ otherwise, where $C(F)$ is easily computable. Thus any approximation algorithm for the single string MLM for this class with guaranteed approximation ratio of $2^{|w|^{1-\alpha}}$ can be used to solve 3-SAT, and hence the problem is NP-hard. The rough idea of the reduction is as follows: Let us imagine that an agent (the hypothetical approximation algorithm) is attempting to find a stochastic matrix which assigns an approximately maximum probability to the string. The string w is conceptually divided into two parts: $w = w_a w_b$. We design the first half of the string in such a way that if the agent is to maximize the probability on it, it will have to 'lean towards' one of 2^n *deterministic* stochastic matrices of a particular kind, which we call '*canonical stochastic matrices*.' These correspond to the 2^n truth assignments for the n variables in F. More precisely, we define a notion of distance between stochastic matrices and show that if the matrix in question is Δ-far from the closest canonical stochastic matrix, then the probability assigned on the first half is less than the optimal by roughly a factor of $(1 - \Delta)^{|w_a|}$. Now the second half of the string 'tests' whether any of these canonical matrices corresponds to a satisfying assignment for the formula F. For any canonical stochastic matrix, or one that is Δ-close to one, the generation probability assigned on the second half of the string will be high if the corresponding truth assignment satisfies F and otherwise will be less by roughly a multiplicative factor of $\Delta^{|w_b|}$. Thus the agent faces the following dilemma: (i) If it tries to be near-optimal on the first half and chooses a small value of Δ, that is, its stochastic matrix is in fact close to one of the canonical matrices, then it would have to in effect solve 3-SAT to determine an approximately maximum generation probability assignable on the second part. (ii) If it tries to avoid solving 3-SAT on the second part and chooses a large enough value of Δ, that is, its stochastic matrix is sufficiently far from any of the canonical matrices, then it loses so much probability on the first half that it cannot guarantee an approximately optimal generation probability.

Proof Sketch of Theorem 5.1

We now give a fuller proof sketch, defining key notions used in the proof in doing so. For each $\alpha > 0$, we exhibit a transformation ω_α mapping any 3-CNF-formula F to a string over an alphabet that depends on the number of variables, for which a gap of factor $K = 2^{|w_\alpha(F)|^{1-\alpha}}$ is forced in the optimal solution depending on the satisfiability of the formula. Let n be the number of variables in F, and s the number of clauses in F. We let Σ_n denote the alphabet used for $\omega_\alpha(F)$, and $\mathcal{M}_n$ the set of all 2-state stochastic matrices over Σ_n. We begin by describing the alphabet Σ_n. The letters in Σ_n can be classified into the following categories: *global control letters* a and b, *fixed function letters* f and e, and for each variable x_i, *literal letters* $x_i\ \bar{x}_i$, *control letters* c_i, d_i, *and a dummy letter* v_i. The literal letters directly correspond to the literals in the formula F, whereas the remaining letters play a support role. We refer to the two states as state 0 and state 1, and let S denote the state set $\{0, 1\}$. We now give the string $\psi_\alpha(F)$, or w for short.

$$w = w_0 w_1 w_2 w_3 w_4 w_5 \tag{5.1}$$

$$w_0 = (ab)^{k_0}$$

$$w_1 = (afabfb)^{k_1}$$

$$w_2 = \prod_{i=1}^{n} v_i (c_i d_i)^{k_2}\ v_i$$

$$w_3 = \prod_{i=1}^{n} (c_i f c_i d_i f d_i)^{k_3}$$

$$w_4 = \prod_{i=1}^{n} (a x_i b x_i c_i \bar{x}_i d_i \bar{x}_i b e a e a b)^{k_4}$$

$$w_5 = \prod_{j=1}^{s} (ab(l_{j,1} l_{j,2} l_{j,3})b)^{k_5}$$

In the above, we let $l_{j,k}$ denote the k-th literal in the j-th clause of F, and the k_i are the integers defined as follows.

$$k_5 = \log K$$

$$k_4 = s 2^{11}(\log K + |\ w_5\ |)$$

$$k_3 = s2^{11}(\log K + | w_4 w_5 |)$$

$$k_2 = s2^{11}(\log K + | w_3 w_4 w_5 |)$$

$$k_1 = s2^{11}(\log K + | w_2 w_3 w_4 w_5 |)$$

$$k_0 = s2^{11}(\log K + | w_1 w_2 w_3 w_4 w_5 |)$$

The k_i were chosen as moderately growing functions of K, the intended *gap*, so that $|w| \leq p(n, s) \log K$ for some polynomial p. If $\alpha > 1$ then reset it to 1. Now ω_α is obtained by setting $\log K = p(n, s)^{1-\alpha/\alpha}$ which makes $|w|$ less than or equal to $p(n, s)^{1/\alpha}$ and hence $\log K \geq | w |^{1-\alpha}$ as desired. We may assume without loss of generality that $\log K$ as set above is an integer because if $p(n, s)^{1-\alpha/\alpha}$ was not an integer then by at most halving α one could find a smaller α' such that $p(n, s)^{1-\alpha'/\alpha'}$ is an integer.

To explain the intent of the transformation given above, we need to introduce some terminology concerning stochastic matrices. A *deterministic* stochastic matrix is a stochastic matrix in which for each state i and each letter z, there is at most one transition with a positive probability out of i labeled with z. Thus any deterministic stochastic matrix M, induces for each letter z a (possibly partial) transition function from states to states. Since for this proof the number of states is 2, these transition functions are Boolean. If the letter has transitions out of each of the two states, then the associated function must be one of the four possible total boolean functions over one variable. Borrowing Angluin's terminology (Angluin, 1989), these are: 0-reset (**0**), 1-reset (**1**), *identity* (**id**) and *flip* (**flip**), and are defined as follows: (i) $\mathbf{0}(0) = \mathbf{0}(1) = 0$, (ii) $\mathbf{1}(0) = \mathbf{1}(1) = 1$, (iii) $\mathbf{id}(0) = 0$, $\mathbf{id}(1) = 1$, and (iv) $\mathbf{flip}(0) = 1$, $\mathbf{flip}(1) = 0$. (See Figure 2.) With a letter that has a transition out

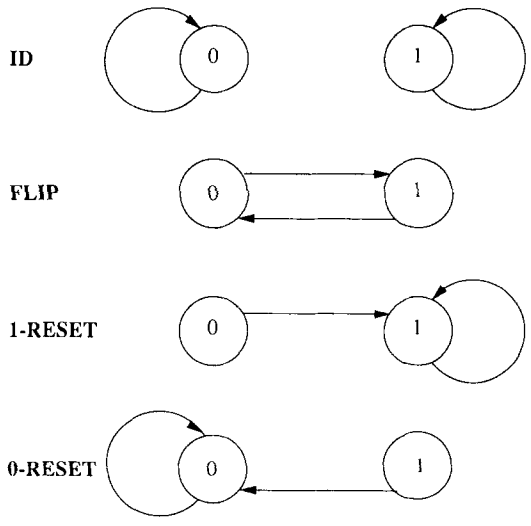

Figure 2. The four boolean functions on two variables.

of only one of the two states we associate one of the four possible *partial* one-variable boolean functions: (v) $0 \to 0$, (vi) $0 \to 1$, (vii) $1 \to 0$, and (viii) $1 \to 1$. We refer to letters with a total (partial) transition function as *total* (respectively *partial*) letters. If in a stochastic matrix M a letter z is associated with **flip**, for example, we write $z =_M$ **flip**. If a pair of letters x and y are **id** and **flip** respectively, then we write $(x, y) =_M$ (**id**, **flip**). (When M is clear from the context, we will drop M and write $(x, y) =$ (**id**, **flip**).)

As explained in the high level description at the beginning of this section, the idea of our proof is that $w_a = w_0w_1w_2w_3w_4$ forces any near-optimal stochastic matrix to be *close to* one of 2^n many *deterministic* stochastic matrices corresponding to truth assignments, and $w_b = w_5$ distinguishes those corresponding to satisfying assignments from those corresponding to non-satisfying assignments. Below we describe how this is intended to be achieved at a high level, leaving the exact nature of the notion of 'closeness' among stochastic matrices to be specified later. The intended function of the first part w_0 of the string w_a is to force the two 'control letters' a and b to go from 0 to 1, and 1 to 0, respectively or vice-versa. The intuitive reason why $(ab)^{k_0}$ forces these settings is as follows: Any stochastic matrix which has two transitions for either a or b will lose probability, and hence there can be only one transition for each of a and b. Furthermore, to be able to generate $(ab)^{k_0}$, the single transition for a and the single transition for b must form a cycle. So, we must have (i) $(a, b) = (0 \to 0, 0 \to 0)$, (ii) $(a, b) = (1 \to 1, 1 \to 1)$, (iii) $(a, b) = (1 \to 0, 0 \to 1)$, or (iv) $(a, b) = (0 \to 1, 1 \to 0)$. But because the length of w_0 is almost the entire length of w, approximately optimal stochastic matrices must let these two transitions have *very* large probabilities (close to one). This is impossible if both of these transitions are out of the same state, so the option (i) and (ii) are eliminated, leaving (iii) and (iv). We assume without loss of generality that we have (iv), that is, $(a, b) = (0 \to 1, 1 \to 0)$. It is easy to see that with these particular settings of a and b, w_1 forces f to be a **flip**. w_2 performs the analogous function for each (c_i, d_i) pair as w_0 did for (a, b), but since w_2 is *not* the overwhelming majority (and there are n such pairs), at this point all four (i–iv) options for a cycle are possible for each (c_i, d_i) pair. w_3 uses f, which has been set to **flip** by w_1, to eliminate (i) and (ii), and forces each (c_i, d_i) pair to be either $(1 \to 0, 0 \to 1)$ or $(0 \to 1, 1 \to 0)$. The crucial observation is that for each i, the direction of (c_i, d_i) *in relationship to* the direction of (a, b) is left unspecified. Utilizing this degree of freedom, w_4 sets the literal letters $x_i, \bar{x}_i$ in a particular way: For each i, $(x_i, \bar{x}_i)$ is forced to be either (**1**, **id**) (see Figure 3) or (**id**, **1**) (see Figure 4), corresponding respectively to the assignment of 'true' and 'false' to the variable x_i. In this way, stochastic matrices assigning a near optimal generation probability on $w_0w_1w_2w_3w_4$ are forced to be *close to* one of these deterministic stochastic matrices, which we called earlier 'canonical stochastic matrices.'

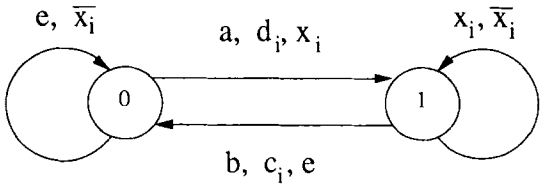

Figure 3. The deterministic automaton corresponding to 'True.'

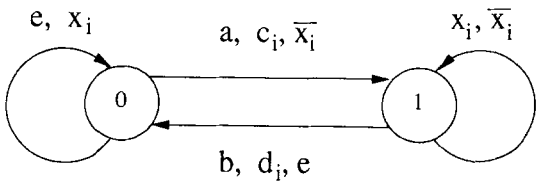

Figure 4. The deterministic automaton corresponding to 'False.'

Observe that there are 2^n of these, corresponding to the 2^n truth assignments on n variables. Finally, w_5 is designed so that any canonical matrix 'satisfying' F will assign it a probability exceeding some bound, whereas any canonical matrix not satisfying F will assign it probability zero. Here we use a trick related to that used by Angluin in Angluin (1989). For a given clause C_j in F, $(x_1, \bar{x}_4, x_2)$ for example, $w_{4,j}$ is $ab(x_1\bar{x}_4x_2)b$. Since each $w_{4,j}$ is preceded by an 'ab' and followed by a 'b,' $(x_1, \bar{x}_4, x_2)$, is forced to map 0 to 1. Now the crucial observation is that if all three letters x_1, $\bar{x}_4$, and x_2 are set **id** (or the corresponding truth assignment assigns all three literals 'false') then so is $(x_1, \bar{x}_4, x_2)$, and hence it must map 0 to 0. Since for any non-satisfying truth assignment there is a clause not satisfied by it, any canonical matrix corresponding to a non-satisfying assignment assigns the string w probability zero. Hence if F is unsatisfiable then any canonical matrix must assign zero probability on $w = \omega_\alpha(F)$.

Proof of Theorem 5.1

Now we make our argument precise. We begin by defining the notion of canonical (deterministic) stochastic matrices.

Definition 5.1 (Canonical Stochastic Matrices). *Let T_n denote the set of truth assignments on n variables, each mapping $\{X_i \mid 1 \le i \le n\}$ to $\{True, False\}$. For each $\tau \in T_n$, we define the 'canonical stochastic matrix for τ,' written M_τ as follows.*

1. *M_τ is deterministic.*
2. *$(a, b) =_{M_\tau} (0 \to 1, 1 \to 0)$, $f =_{M_\tau}$ **flip**, and $e =_{M_\tau}$ **0**.*
3. *For each variable X_i,*

$$
\begin{array}{lll}
(c_i, d_i) & =_{M_\tau} (1 \to 0, 0 \to 1) & \text{if } \tau(X_i) = True. \\
 & =_{M_\tau} (0 \to 1, 1 \to 0) & \text{if } \tau(X_i) = False. \\
v_i & =_{M_\tau} \textbf{flip} & \text{if } \tau(X_i) = True. \\
 & =_{M_\tau} \textbf{id} & \text{if } \tau(X_i) = False. \\
(x_i, \bar{x}_i) & =_{M_\tau} (\textbf{1}, \textbf{id}) & \text{if } \tau(X_i) = True. \\
 & =_{M_\tau} (\textbf{id}, \textbf{1}) & \text{if } \tau(X_i) = False.
\end{array}
$$

4. *Each non-zero transition probability in M_τ is either equal to or twice the* frequency *in w of the letter z labeling the transition where this frequency is defined as the number of occurrences of z in w, written $\#(z, w)$, divided by $|w|$.*
   (i) For any partial letter z, $M_\tau(i, j, z) = 2\#(z,w)/|w|$ if and only if $M_\tau(i, j, z) \neq 0$.
   (ii) For any total letter z, $M_\tau(i, j, z) = \#(z,w)/|w|$ if and only if $M_\tau(i, j, z) \neq 0$.
   (Notice that a and b, and all c_i and d_i are partial letters and all other letters are total.)

We still need to show that M_τ is stochastic. For convenience we will defer this proof till (the paragraph following (5.5)), after we have developed some more notation for stochastic matrices. Note that in the above definition we have arbitrarily chosen a to go from 0 to 1 and b to go from 1 to 0. Given a canonical stochastic matrix M_τ, we let M_τ^r denote its *mirror image*, obtained by reversing the states 0 and 1 in M_τ. $M_\tau(w)$ can be calculated easily because canonical stochastic matrices are deterministics: $M_\tau(w)$ is zero, if τ does not satisfy F, and otherwise is just the product of the probabilities assigned on all the transitions occurring in *the* unique path for w that is assigned a positive probability by M_τ. If we define $C(w)$ as follows, regardless of whether F is satisfiable or not,

$$C(w) = \prod_{z \text{ partial}} \left(2 \frac{\#(z, w)}{|w|} \right)^{\#(z,w)} \cdot \prod_{z \text{ total}} \left(\frac{\#(z, w)}{|w|} \right)^{\#(z,w)} \tag{5.2}$$

then we have $M_\tau(w) = C(w)$ just in case τ satisfies F, and $M_\tau(w) = 0$ otherwise. We are now ready to state the key lemma in the proof of Theorem 5.1.

Lemma 5.1 *Let ω_α be as defined in (5.1), and let $M^*(w)$ denote the maximum probability assignable on w by any 2-state PA, i.e. $M^*(w) = \max_{M \in \mathcal{M}_n} M(w)$. Then for any CNF formula F and $K = 2^{|w|^{1-\alpha}}$, we have:*

1. If F is satisfiable, then $M^*(\omega_\alpha(F)) \geq C(\omega_\alpha(F))$.
2. If F is unsatisfiable, then $M^*(\omega_\alpha(F)) < C(\omega_\alpha(F))/K$.

Proof of Lemma 5.1

The proof of part 1 is immediate: If F is satisfiable then let τ satisfy F and we have $M_\tau(w) = C(\omega_\alpha(F))$. The other direction is more involved and requires more definitions. The key is to find a useful way of quantifying the distance δ between an *arbitrary* stochastic matrix M and canonical stochastic matrices such that the generation probability assigned on w by M can be shown to degrade rapidly as a function of $min_{\tau \in T_n} \delta(M, M_\tau)$. We can then use it to carry out the dilemma argument described in the proof sketch. We need some preliminary definitions.

Recall that for a given stochastic matrix M, $M(i,j,z)$ denotes the transition probability from state i to j labeled with letter z. We introduce the following notation for sums of transition probabilities of various forms:

$$M(*, *, z) = \prod_{i,j \in S} M(i, j, z)$$

$$M(i, *, z) = \prod_{j \in S} M(i, j, z)$$

$$M(*, j, z) = \sum_{i \in S} M(i, j, z)$$

We call $M(i, *, z)$ the *out-share* of z at i in M, $M(*, j, z)$ the *in-share of* z at j in M, and $M(*, *, z)$ the *share* of z in M. We partition our alphabet Σ_n into *letter groups*, in which the pairs of partial letters (a, b) and (c_i, d_i) are grouped together and each total letter forms a group with itself as the only letter.[9] Let Γ_n denote the set of these letter groups, that is, $\Gamma_n = \{\{a, b\}\} \cup \{\{c_i, d_i\} : i = 1, \ldots, n\} \cup \{\{z\} : z \text{ total}\}$. We then define the *share of a letter group* H in a stochastic matrix M, written $M(*, *, H)$, to be the total sum of the probabilities of all transitions in M labeled with a letter in H. That is,

$$M(*, *, H) = \sum_{i,j \in S, z \in H} M(i, j, z)$$

The *out-share* and *in-share* of a *letter group* are defined analogously to those of a letter and $\#(H, w)$ is defined as $\Sigma_{z \in H} \#(z, w)$. Recall that in a canonical stochastic matrix, there is one transition out of each state for each total letter whose probability is the frequency of that letter, and there is exactly one transition for each partial letter, whose probability is twice the frequency of that letter. This implies that the share of each letter group in M_τ is set according to twice the total frequency of the letters from that letter group:

$$\forall H \in \Gamma_n \; M_\tau(*, *, H) = 2 \frac{\#(H, w)}{|w|} \tag{5.3}$$

For example, for the letter group consisting of a single total letter f, we have the following from the definition of M_τ.

$$M_\tau(*, *, f) = M_\tau(0, *, f) + M_\tau(1, *, f) = 2 \frac{\#(f, w)}{|w|} \tag{5.4}$$

For a letter group consisting of two partial letters, for example (a, b) we can derive the following also from the definition of M_τ.

$$M_\tau(*, *, (a, b)) = M_\tau(*, *, a) + M_\tau(*, *, b) = 2 \frac{2\#(a, w)}{|w|} = 2 \frac{\#((a, b), w)}{|w|} \tag{5.5}$$

For each letter group the share is split evenly to the two states, and hence $M_\tau(0, *, H) = M_\tau(1, *, H) = \#(H, w)/|w|$. By summing over all letter groups it follows that $M_\tau(0, *, *) = M_\tau(1, *, *) = 1$ and thus M_τ of Definition 5.1 is stochastic. The generation probability $C(w)$ can now be rewritten as follows, by plugging (5.3) into (5.2).

$$C(w) = \prod_{H \in \Gamma_n} \left(\frac{\#(H, w)}{|w|} \right)^{\#(H,w)} = \prod_{H \in \Gamma_n} \left(\frac{M_\tau(*, *, H)}{2} \right)^{\#(H,w)}, \quad (5.6)$$

Note that we can straightforwardly generalize $C(w)$ to an arbitrary substring u of w: We define $C(u)$ by replacing each occurrence of $\#(H, w)$ by $\#(H, u)$:

$$C(u) = \prod_{H \in \Gamma_n} \left(\frac{M_\tau(*, *, H)}{2} \right)^{\#(H,u)} \quad (5.7)$$

It can be shown that $C(u)$ is the probability that any *satisfying* canonical stochastic matrix assigns on u, but in this proof we only use the above definition for $C(u)$. We also introduce a version of $C(u)$ which is *relativized* to the letter group shares of an arbitrary stochastic matrix M. That is, for an arbitrary string u we define the quantity $C[M](u)$ as follows:

$$C[M](u) = \prod_{H \in \Gamma_n} \left(\frac{M(*, *, H)}{2} \right)^{\#(H,u)} \quad (5.8)$$

Again there is an interpretation of $C[M](u)$ as the probability assigned on u, when τ satisfies F, by a variant $M_\tau[M]$ of the canonical stochastic matrix M_τ which is defined exactly analogously to M_τ except the letter group shares of $M_\tau[M]$ are the same as those of M. Note that $C[M](uv) = C[M](u)C[M](v)$.

We now formalize the notion of distance between an arbitrary stochastic matrix M and any canonical stochastic matrix M_τ. The *H-leak* of an arbitrary stochastic matrix M with respect to M_τ at state i, written $\lambda_i^H(M, M_\tau)$, is the fraction of H's share out of i which is assigned by M to transitions assigned zero probability by M_τ:

$$\lambda_i^H(M, M_\tau) = \sum_{z \in H, j \in S, M_\tau(i,j,z)=0} \frac{M(i, j, z)}{M(i, *, H)} \quad (5.9)$$

Also,

$$\lambda^H(M, M_\tau) = max_{i \in S} \lambda_i^H(M, M_\tau) \quad (5.10)$$

We then define the *leak* of M with respect to M_τ as the maximum $\lambda_i^H(M, M_\tau)$ over all states and letter groups.

$$\lambda(M, M_\tau) = max_{H \in \Gamma_n} \lambda^H(M, M_\tau) \quad (5.11)$$

The *skew*, written $\nu(M)$, is twice the maximum deviation of the ratio $M(0, *, H)/M(*, *, H)$ from a half, where the maximum is over all letter groups H.

$$\nu^H(M) = 2 \mid \frac{M(0, *, H)}{M(*, *, H)} - \frac{1}{2} \mid \text{ and}$$

$$\nu(M) = max_{H \in \Gamma_n} \nu^H(M) \tag{5.12}$$

Note that since $0 \leq M(0, *, H)/M(*, *, H) \leq 1$, we have, for an arbitrary letter group H:

$$0 \leq \nu^H(M) \leq 1 \tag{5.13}$$

Recall that $M_\tau(0, *, H)/M_\tau(*, *, H) = 1/2$ and thus for motivational purposes the skew of M for a letter group H can be expressed as the ratio at which the ratio $M(0, *, H)/M(*, *, H)$ deviates from $M_\tau(0, *, H)/M_\tau(*, *, H)$:

$$\nu^H(M) = \frac{\left| \frac{M(0, *, H)}{M(*, *, H)} - \frac{M_\tau(0, *, H)}{M_\tau(*, *, H)} \right|}{\frac{M_\tau(0, *, H)}{M_\tau(*, *, H)}}$$

We then define the *distortion* of M with respect to any deterministic stochastic matrix M_τ, written $\delta(M, M_\tau)$, as follows:

$$\delta(M, M_\tau) = max \{\lambda(M, M_\tau), \nu(M)^2\} \tag{5.14}$$

The *distortion* of M is then defined as the distortion of M with respect to the *closest* canonical stochastic matrix, including their 'mirror images.'

$$\delta(M) = min_{\tau \in T_n} min \{\delta(M, M_\tau), \delta(M, M_\tau^r)\}$$

For each of *leak*, *skew* and *distortion*, *we define the restriction* of it to an arbitrary subset Ψ of the letter groups by maximizing over Ψ instead of Γ_n in the above definitions. We use readable names such as (a, b) and (a, b, c, d) to refer to subsets of Γ_n such as $\{\{a, b\}\}$ and $\{\{a, b\}\} \cup \{\{c_i, d_i\}\} : i = 1, \ldots, n\}$, respectively. We then let symbols such as $\lambda^{(a,b)}$, $\nu^{(f)}$, and $\delta^{(a,b,c,d)}$ denote the corresponding restrictions of λ, ν and δ. Note that the only information in M_τ that was used to define $\lambda(M, M_\tau)$ is the transition function associated with M_τ, and to define $\lambda^H(M, M_\tau)$, only the restriction of it to H. Thus, for any transition function $f: S \times \Sigma_n \to S$, restricted to a letter group H, we define $\lambda^H(M, f)$ to be $max_{i \in S, H \in \Gamma_n} \Sigma_{z \ni H, j \in S, f(i,z) \neq j} M(i, j, z)/M(i, *, H)$, and define $\delta^H(M, f)$ accordingly. For readability, we use notation such as $\lambda^{(a,b)}(M, \{a : 0 \to 1, b : 1 \to 0\})$ and $\delta^{(a,b,c_i,d_i)}(M, \{a : 0 \to 1, b : 1 \to 0, c_i : 0 \to 1, d_i : 0 \to 0\})$.

The notion of distortion just defined quantifies how bad an arbitrary stochastic matrix is in comparison to canonical stochastic matrices which are *near optimal*, from the point of view of maximizing the generation probability on the *unique* path on the substring $w_0w_1w_2w_3w_4$, which is assigned a positive generation probability by canonical stochastic

matrices. Any leak causes a loss of probability on this path as is obvious from its definition. Any skew also causes a loss of probability on the path, essentially because the path in question for the string $w_0w_1w_2w_3w_4$ is almost perfectly symmetric in the following sense: For each letter group (except the dummy letters v_i) there are exactly the same number of transitions out of state 0 as there are out of state 1. We demonstrate this via an example.

Example 5.1. *Let M be an arbitrary stochastic matrix, and M_τ a canonical stochastic matrix. Consider the string ab and the path for ab which is assigned a positive probability by M_t, namely, $0 \to 1 \to 0$. We will show that the probability assigned by M on this path, $M(0, 1, a) \cdot M(1, 0, b)$ is at most $(1 - \delta^{(a,b)}(M))(M(*, *,(a, b))/2)^2$. Notice that $(M(*, *, \{a, b\})/2)^2$ is the maximum probability assignable on ab by any matrix $\hat{M}$ such that $\hat{M}(*, *, (a, b)) = M(*, *, (a, b))$. To show the above inequality first observe that*

$$M(0, 1, a) \cdot M(1, 0, b) \leq (1 - \lambda_0^{(a,b)}(M, M_\tau))M(0, *, (a, b)) \cdot (1 - \lambda_1^{(a,b)}(M, M_\tau))M(1, *, (a, b))$$
$$\leq (1 - max\{\lambda_0^{(a,b)}(M, M_\tau), \lambda_1^{(a,b)}(M, M_\tau)\})M(0, *, (a, b))M(1, *, (a, b))$$

By (5.10) the above maximization clause can be replaced by $\lambda^{(a,b)}(M, M_\tau)$. Also assume without loss of generality that $M(0, *, (a, b)) \geq M(1, *, (a, b))$. From the definition of $\nu^{(a,b)}$ in (5.12), it follows that $M(0, *, (a, b)) = (1 + \nu^{(a,b)})M(*, *, (a, b))/2$ which implies that $M(1, *, (a, b)) = (1 - \nu^{(a,b)})M(*, *, (a, b))/2$ and we continue as follows:

$$M(0, 1, a) \cdot M(1, 0, b) \leq (1 - \lambda^{(a,b)}(M, M_\tau))(1 + \nu^{(a,b)}(M))(1 - \nu^{(a,b)}(M)) \left(\frac{M(*, *, (a,b))}{2}\right)^2$$
$$\leq (1 - \lambda^{(a,b)}(M, M_\tau))(1 - \nu^{(a,b)}(M^2)) \cdot \left(\frac{M(*, *, (a,b))}{2}\right)^2$$
$$\leq (1 - max\{\lambda^{(a,b)}(M, M_\tau), \nu^{(a,b)}(M)^2\}) \cdot \left(\frac{M(*, *, (a,b))}{2}\right)^2$$
$$= (1 - \delta^{(a,b)}(M, M_\tau)) \cdot \left(\frac{M(*, *, (a,b))}{2}\right)^2 \text{ by (5.14)}$$

The following technical lemmas are useful for proving Lemma 5.1, part 2. Their proofs are deferred to Appendix A.

Lemma 5.2. *For an arbitrary stochastic matrix $M \in \mathcal{M}_n$, we have:*

$$C(w) \geq C[M](w)$$

Lemma 5.3. *For arbitrary $M \in \mathcal{M}_n$ and $x \in \Sigma_n^*$ we have:*

1. $M(x) \leq \Pi_{z \in \Sigma_n} (max\{M(0, *, z), M(1, *, z)\})^{\#(z,x)}$.
2. $M(x) \leq \Pi_{z \in \Sigma_n} (max\{M(*, 0, z), M(*, 1, z)\})^{\#(z,x)}$.

Lemma 5.4. *For an arbitrary stochastic matrix $M \in \mathcal{M}_n$ and an arbitrary string $x \in \Sigma_n^*$, we have:*

$$M(x) \leq 2^{|x|} C[M](x)$$

Lemma 5.5. *Let $M \in \mathcal{M}_n$ be an arbitrary stochastic matrix. Let $\delta(M) = \Delta$ and let M_τ be a canonical stochastic matrix to which M is closest, namely $\delta(M, M_\tau) = \Delta$. If $M_\tau(x) = 0$ then we have for an arbitrary string x:*

$$M(x) \leq |x| \cdot \Delta(1 + \sqrt{\Delta})^{|x|} C[M](x)$$

Using the above technical lemmas, we are now ready to prove Lemma 5.1, part 2. The following lemma summarizes the key steps of this proof. Its proof also uses the above technical lemmas and is relegated to Appendix B.

Lemma 5.6. *For an arbitrary stochastic matrix $M \in \mathcal{M}_n$ all of the following hold.*

1. *If $M(w) \geq 1/K\ C(w)$, then $\delta^{(a,b)}(M) \leq 2^{-10}/s$.*
2. *$M(w_0) \leq C[M](w_0)$.*
3. *If $\delta^{(a,b)}(M) \leq 2^{-10}/s$ then $M(w_1) < (1 - \delta^{(a,b,f)}(M)/2)^{k_1} C[M](w_1)$.*
4. *For each $i \leq n$ let $\Delta_i = min\{\delta^{(c_i,d_i)}(M, \{c_i : 0 \to 0, d_i : 0 \to 0\}), \delta^{(c_i,d_i)}(M, \{c_i : 0 \to 1, d_i : 1 \to 0\}), \delta^{(c_i,d_i)}(M, \{c_i : 1 \to 0, d_i : 0 \to 1\}), \delta^{(c_i,d_i)}(M, \{c_i : 1 \to 1, d_i : 1 \to 1\})\}$. and let $\Delta = max_{i \leq n} \Delta_i$. Then, if $M(w) \geq C(w)/K$ then $\Delta \leq 2^{-10}/s$.*
5. *If $\Delta \leq 2^{-10}/s$, where Δ is as defined above, and $\delta^{(f)}(M) \leq 2^{-10}/s$, then $M(w_3) < (1 - \delta^{(c,d,f)}(M)/2)^{k_3} C[M](w_3)$.*
6. *If $\delta^{(a,b,c,d,f)}(M) \leq 2^{-10}/s$ then $M(w_4) < (1 - \delta^{(a,b,c,d,f,x,\bar{x},e)}(M)/2)^{k_4} C[M](w_4)$.*
7. *If $\delta^{(a,b,c,d,f,x,\bar{x},e)}(M) \leq 2^{-10}/s$ and F is unsatisfiable, then $M(w_5) < (1/2)^{k_5} C[M](w_5)$.*

Proof of Lemma 5.1, part 2, given Lemmas 5.2, 5.4 and 5.6. Assume that $M^*(w) \geq C(w)/K$, and let a particular M witness this fact. We will show that this will imply that F is satisfiable. First, the following follows immediately from Lemma 5.6, part 1.

$$\delta^{(a,b)}(M) \leq \frac{1}{s2^{10}}. \tag{5.15}$$

Next, suppose for contradiction that $\delta^{(f)}(M) > 2^{-10}/s$. Then, $\delta^{(a,b,f)}(M) = \delta^{(f)}(M)$, since $\delta^{(f)}(M) > \delta^{(a,b)}(M)$. It follows from Lemma 5.6, part 3:

$$M(w_1) \leq \left(1 - \frac{\delta^{(f)}(M)}{2}\right)^{k_1} C[M](w_1)$$

Substituting $k_1 = s2^{11}(\log K + | w_2 w_3 w_4 w_5 |)$ into the above gives us:

$$M(w_1) \leq \left(1 - \frac{\delta^{(f)}(M)}{2}\right)^{s2^{11}(\log K + |w_2w_3w_4w_5|)} \cdot C[M](w_1)$$

$$\leq e^{-\delta^{(f)}(M)/2 \cdot s2^{11}(\log K + |w_2w_3w_4w_5|)} \cdot C[M](w_1)$$

$$< \frac{1}{K}\left(\frac{1}{2}\right)^{|w_2w_3w_4w_5|} \cdot C[M](w_1) \tag{5.16}$$

Lemma 5.4 provides us with an upper bound on how large $M(w_2w_3w_4w_5)$ could possibly be in comparison to $C[M](w_2w_3w_4w_5)$, that is,

$$M(w_1w_2w_3w_4w_5) \leq 2^{|w_1w_2w_3w_4w_5|} C[M](w_1w_2w_3w_4w_5)$$

We can use this bound and (5.16) to bound $M(w)$ above by $1/K\ C(w)$ as follows:

$$M(w) \leq M(w_0)M(w_1)M(w_2w_3w_4w_5)$$

$$\leq C[M](w_0)M(w_1)M(w_2w_3w_4w_5), \text{ by Lemma 5.6, part 2}$$

$$< C[M](w_0) \cdot \frac{1}{K}\left(\frac{1}{2}\right)^{|w_2w_3w_4w_5|} C[M](w_1) \cdot M(w_2w_3w_4w_5), \text{ by (5.16)}$$

$$= \frac{1}{K} \cdot C[M](w_0) \cdot C[M](w_1) \cdot \left(\frac{1}{2}\right)^{|w_2w_3w_4w_5|} M(w_2w_3w_4w_5), \text{ by rearranging terms}$$

$$< \frac{1}{K} \cdot C[M](w_0) \cdot C[M](w_1) \cdot C[M](w_2w_3w_4w_5), \text{ by Lemma 5.4}$$

$$< \frac{1}{K} \cdot C[M](w), \text{ by definition of } C[M](w)$$

$$\leq \frac{1}{K} \cdot C(w), \text{ by Lemma 5.2} \tag{5.17}$$

This contradicts our assumption, and hence together with (5.15), we conclude:

$$\delta^{(a,b,f)}(M) = max\{\delta^{(a,b)}(M), \delta^{(f)}(M)\} \leq \frac{1}{s2^{10}}. \tag{5.18}$$

Since by assumption, $M(w) \geq C(w)/K$, Δ as defined in the statement of Lemma 5.6, part 4, does not exceed $2^{-10}/s$. Suppose now for contradiction that $\delta^{(c,d)}(M) > 2^{-10}/s$. Then, since $\delta^{(a,b)}(M)$, Δ and $\delta^{(f)}(M)$ are now known not to exceed $2^{-10}/s$, it follows from Lemma 5.6, parts 2 through 5, and a similar argument as before, that this would imply $M(w) < 1/K\ C(w)$. Thus, we conclude:

$$\delta^{(a,b,c,d,f)}(M) \leq \frac{1}{s2^{10}} \tag{5.19}$$

Given the above, it follows from Lemma 5.6, parts 2 through 6, by an analogous argument as before, that the distortion of M with respect to all letters is small, that is,

$$\delta(M) \leq \frac{1}{s2^{10}} \tag{5.20}$$

Finally, suppose that F *is unsatisfiable.* Then by Lemma 5.6, part 7, we must have

$$M(w_5) < \left(\frac{1}{2}\right)^{k_5} \cdot C[M](w_5)$$

Now again using Lemma 5.6, parts 2 through 6, and the fact that $k_5 = \log K$, we can show that $M(w) < 1/K\ C(w)$, contradicting our assumption. Hence we conclude that F must be satisfiable.

End of proof of Lemma 5.1 and proof of Theorem 5.1

Proof of Theorem 5.2. First note that 1-state PAs are *deterministic.* Thus, by Corollary 4.2, it is clear that the MLM problem for the 1-state constraints is solvable in polynomial time. We then show that if we use the optimal 1-state PA for the MLM problem for s-state null constraints, then it achieves the guaranteed approximation factor of $s^{|w|}$. The optimal 1-state PA sets the transition probability of each letter proportionally to the frequency of that letter in the input string w. More precisely, the optimal matrix, denoted M^*, is defined as follows: For each letter $z \in \Sigma$, we set

$$M^*(0, 0, z) = \frac{\#(z, w)}{|w|}$$

where we let 0 be the unique state in the 1-state PA. Hence, the probability M^* assigns on w is easily computed as follows:

$$M^*(w) = \prod_{z \in \Sigma} \left(\frac{\#(z, w)}{|w|}\right)^{\#(z,w)}$$

Now, using Lemma 5.3, we can compute the following upper bound on $M(w)$ for any s-state stochastic matrix M.

$$M(w) \leq \prod_{z \in \Sigma} (max_{i \in S} M(i, *, z))^{\#(z,w)} \tag{5.21}$$

Here we have the constraint that $\Sigma_{z \in \Sigma}\ max_{i \in S} M(i, *, z) \leq s$. Hence, by Lemma 4.2, we have:

$$M(w) \leq \prod_{z \in \Sigma} \left(s \frac{\#(z, w)}{|w|} \right)^{\#(z,w)}$$

$$s^{|w|} \prod_{z \in \Sigma} \left(\frac{\#(z, w)}{|w|} \right)^{\#(z,w)}$$

$$= s^{|w|} M^*(w)$$

End of proof of Theorem 5.2

6. Application to training hidden Markov models

Hidden Markov models are used extensively as probabilistic models for generation of speech signals for the purpose of speech recognition. See Levinson, Rabiner and Sandhi (1983) for an excellent tutorial on this material. Hidden Markov models are defined similarly to probabilistic automata, except that the generation of letters is associated with the *states* rather than with the *transitions*. We briefly review the definition below.

Definition 6.1 (Hidden Markov Models(HMM)). *A hidden Markov model P is a quintuple $\langle S_P, \Sigma_P, \pi_P, M_P, L_P \rangle$ where S_P is a finite set of states, Σ_P is a finite alphabet, $\pi_P : S_P \to [0, 1]$ is the initial probability distribution over S_P, $M_P : S_P \times S_P \to [0, 1]$ is a stochastic matrix, and $L_P : S_P \times \Sigma_P \to [0, 1]$ specifies the letter generation distribution at each state, i.e.*

$$\sum_{i \in S_P} \pi_P(i) = 1 \text{ and } \forall i \in S_P \sum_{j \in S_P} M_P(i, j) = 1 \text{ and } \forall i \in S_P \sum_{a \in \Sigma_P} L_P(i, a) = 1 \quad (6.1)$$

For any string $w = w_1 \ldots w_n \in \Sigma_P^*$, the generation probability assigned on it by $P = \langle S_P, \Sigma_P, \pi_P, M_P, L_P \rangle$ is computed as follows.

$$P(w_1 \ldots w_n) = \sum_{\langle i_0, \ldots, i_n \rangle \in S_P^{n+1}} \pi_P(i_0) \cdot \prod_{j=0}^{n-1} L_P(i_j, w_{j+1}) M_P(i_j, i_{j+1}) \quad (6.2)$$

As before, for any given example length n, P defines a probability distribution over Σ_P^n.

Any hidden Markov model can be simulated by a probabilistic automaton of the same number of states. For an arbitrary hidden Markov model $P = \langle S_P, \Sigma_P, \pi_P, M_P, L_P \rangle$, define a probabilistic automaton $Q = \langle S_Q, \Sigma_Q, \pi_Q, M_Q \rangle$ by letting $S_Q = S_P$, $\Sigma_Q = \Sigma_P$, $\pi_Q = \pi_P$, and

$$\forall i, j \in S_Q \; \forall z \in \Sigma_Q \quad M_Q(i, j, z) = L_P(i, z) \cdot M_P(i, j) \quad (6.3)$$

Then, P and Q define the same distribution over Σ_P^n for any n, as demonstrated below.

$$Q(w_1 \ldots w_n) = \sum_{\langle i_0,\ldots,i_n\rangle \in S_Q^{n+1}} \pi_Q(i_0) \cdot \prod_{j=0}^{n-1} M_Q(i_j, i_{j+1}, w_{j+1}) \text{ by (2.2)}$$

$$= \sum_{\langle i_0,\ldots,i_n\rangle \in S_Q^{n+1}} \pi_P(i_0) \cdot \prod_{j=0}^{n-1} L_P(i_j, w_{j+1}) M_P(i_j, i_{j+1}) \text{ by (6.3)}$$

$$= P(w_1 \ldots w_n) \text{ by (6.2)}$$

Note that the ratio $M_Q(i, j, z')/M_Q(i, j, z')$ equals $L_P(i, z)/L_P(i, z')$ and is independent of j. In other words, the class of s state hidden Markov models is equivalent to the class of s state probabilistic automata satisfying the following condition:

$$\forall i, j, k \in S_P \; \forall z, z' \in \Sigma_P \; \frac{M_Q(i, j, z)}{M_Q(i, j, z')} = \frac{M_Q(i, k, z)}{M_Q(i, k, z')}$$

We define the training problem for hidden Markov models exactly analogously to the training problem for probabilistic automata (c.f. Definition 2.4). The only difference is that the input constraint that an HMM training algorithm receives is an *HMM constraint*, in place of a PA constraint. Here, an HMM constraint is a five-tuple $C = \langle \Sigma, S, I, G, L \rangle$, specifying respectively the finite alphabet, set of states, legal initial states, legal transitions and legal letter generations. Formally,, $I \subseteq S$, $G \subseteq S \times S$ and $L \subseteq S \times \Sigma$. As before, we measure the size of the input constraint by the total number of probability parameters in it, $|I| + |G| + |L|$, and denote this by t. We say that an HMM P satisfies $C = \langle \Sigma, S, I, G, L \rangle$ if an only if $S_P = S$ and $\Sigma_P = \Sigma$ and

$$\begin{aligned} \forall i \notin I, \; \pi_P(i) &= 0 \\ \forall (i, j) \notin G, \; M_P(i, j) &= 0 \\ \forall (i, z) \notin L, \; L_P(i, z) &= 0 \end{aligned} \tag{6.4}$$

For an arbitrary HMM constraint $C = \langle \Sigma, S, I, G, L \rangle$, let $\mathcal{HMM}(C)$ denote the class of HMMs satisfying C. We can think of $\mathcal{HMM}(C)$ as the subset of $[0, 1]^I \times [0, 1]^G \times [0, 1]^L$ satisfying the stochastic condition (6.1). The training problem for a class C of HMM constraints is defined analogously to the training problem for a class of PA constraints (see Definition 2.4). We can show the same sample complexity bound (up to a constant factor) for training a class of HMM constraints as we did for PAs in Theorem 3.1. We only sketch how this is shown, since the proof is almost identical to the proof of Theorem 3.1. We define $\mathcal{BHMM}(C)_m$ by bounding the initial, transition and letter generation probabilities from below by $1/tm$, where t is the size of the input constraint. We define $\mathcal{FHMM}(C)_m$ by quantizing all the probabilities in the same way as before. The rest of the proof is essentially the same. The analogues of inequalities (3.13) and (3.14) for $\mathcal{HMM}(C$, $\mathcal{BHMM}(C)_m$ and $\mathcal{FHMM}(C)_m$ can be shown the same way, noting the fact that the probability assigned on a string of length n by an HMM is a product of $2n + 1$ probabilities, as opposed to $n + 1$ for PAs.

Corollary 6.1. *An arbitrary class of HMM constraints C can be trained with sample complexity $O((n/\epsilon)^2 t \cdot \log^3 nt/\epsilon \cdot \log 1/\delta \cdot \log^2 \log 1/\delta)$, where to is the size of the input constraint.*

Corollary 6.2. *Any finite class of HMM constraints is polynomially trainable.*

7. Concluding remarks

We were able to show that training an arbitrary class of PAs and HMMs can be done with polynomial sample complexity when computational efficiency is ignored. The sample complexity bounds given in Theorem 3.1 for training a class of PAs and in Corollary 6.1 for training a class of HMMs may perhaps be improved significantly. Lower bounds for these training problems should be investigated.

Our method for obtaining sample complexity bounds can be summarized as follows:

(i) We bound the parameter values away from zero to avoid the unboundedness of the Kullback-Leibler divergence.
(ii) We quantize the bounded parameter space to obtain a hypothesis class of a moderate cardinality.
(iii) We show that the resulting quantized hypothesis class 'finely covers' the whole hypothesis class with respect to log loss.
(iv) Finally we apply Hoeffding's inequality on a class of bounded random variables defined in terms of the quantized hypothesis class to obtain an upper bound on the sample complexity for uniform convergence.

It would be interesting to use this method to establish sample complexity bounds for various other parameterized distribution learning problems with respect to the Kullback-Leibler divergence and for such problems with respect to the other measures of distance between distributions mentioned in the introduction.[10] All our sample complexity bounds with respect to the Kullback-Leibler divergence rely on Hoeffding's inequality and thus grow with $1/\epsilon^2$. Is the $1/\epsilon^2$ growth in the sample complexity really necessary?

We showed in Section 6 that s-state HMMs can easily be simulated by s-state PAs. How can HMMs be used to simulate PAs?

There are many open problems related to the hardness results of Section 5. First, we would like to know whether the following decision problem is in **NP.**

Input: two numbers s and a encoded in unary, a string $w \in \Sigma^*$ where $|\Sigma| = a$, a probability $q \in [0, 1]$ encoded in binary.

Question: Does there exist an s-state PA P with alphabet size a such that $P(w) \geq q$?

It would also be interesting to determine the precise computational complexity of various formulations of the approximate MLM problem as a kind of decision problem.

Second, Osamu Watanabe has brought to our attention that the reduction we exhibit in the proof of Theorem 5.1 can probably be modified so as to strengthen our result and show that the approximate single string MLM problem, perhaps with a more strict requirement of approximation, is in fact Δ_2^p-complete.

Third, we would like to know whether the single string MLM problem for the null 2-state HMM constraints (with variable alphabet size) is approximable in polynomial time or can we obtain a similar hardness result for this problem as the one proven in Theorem 5.1 for the single string MLM problem for 2-state PAs with variable alphabet size?

Fourth, can the latter hardness result on the single string MLM problem be strengthened from a factor of $2^{|w|^{1-\alpha}}$ from the optimum (for any $\alpha > 0$) to a factor of $2^{(1-\alpha)\cdot|w|}$ from the optimum (for any $\alpha > 0$)?

Fifth, recall that we have shown in Theorem 4.1 that for an arbitrary class of constraints the approximability of the *sample* MLM problem for it within a factor $2^{m^{1-\alpha}}$ (for any $\alpha > 0$) where m is the sample size, would imply polynomial trainability of the same class. Could this be strengthened so that polynomial approximability within a factor $2^{(1-\alpha)\cdot m}$ (for some $\alpha > 0$) would already imply polynomial trainability of the class in question? Could we perhaps show that polynomial approximability of the *single string* MLM problem within a factor of $2^{|w|^{1-\alpha}}$ or $2^{(1-\alpha)\cdot|w|}$ (for some $\alpha > 0$) would imply polynomial trainability?

Finally, we emphasize that even though the hardness results may be disappointing they can serve as guidance in the search for constructive results. Perhaps the most significant open problem inspired by the results of the present paper is to determine practically relevant classes of PAs and HMMs that are provably polynomially trainable. For example, the class of HMM constraints used in speech recognition (Levinson, Rabiner & Sondhi, 1983) consist of chains of states in which only transitions that go forward in the chain or stay stationary in the chain are legal. What is the lowest sample complexity required for training this important class of HMM constraints? Is this class polynomially trainable or is training this class hard modulo some weak assumption such as **RP** $\neq$ **NP**?

Acknowledgments

Most of this work was done while the first author was at the Department of Computer and Information Sciences, University of California, Santa Cruz, supported by the Office of Naval Reserach under contract number N0014-86-I-0454. Part of this work was done after the first author began employment by NEC Corporation, and while the second author was visiting IIAS-SIS Fujitsu, Limited, Japan.

We thank David Haussler for teaching us many of the tools applied in this paper. In particular, he suggested the idea of bounding transition probabilities from below for showing sample complexity bounds. We thank Dana Angluin for showing us her hardness proof for the 2-state DFA consistency problem (Angluin, 1989) which served as a starting point for the non-approximability result for the single string MLM problem for the 2-state PA constraints. Thanks to Ron Rivest for bringing the single string MLM problem to our attention. Thanks to Nicolo Cesa-Bianchi, Yoav Freund, Phil Long, Aleksandar Milosavljevic, Dirk Van Compernolle, and Osamu Watanabe for fruitful discussions.

Notes

1. HMMs are similar to probabilistic automata, except that outputs in an HMM are associated with the states rather than the transitions, and thus the transitions are unlabeled state to state pairs.
2. The existence of an algorithm for training hidden Markov models which always outputs a *near local optimum* on a given sample is well-known ('Baum-Welch' algorithm (Baum, 1972)) and is used extensively in practice. Note that in applications to speech recognition, the alphabet size is determined by how precise the acoustic signals are quantized. The alphabet size is often in the hundreds.
3. Contrary to what may seem to be the case from the equivalence result described in the previous paragraph (of approximability within a factor of $1 + \epsilon$ and that within $2^{p(n,t)m^{\alpha}}$ for the *sample* MLM problem), we cannot use this equivalence to obtain the non-approximability result whtin a factor of $2^{|w|^{1-\alpha}}$ from that within $1 + \epsilon$, since the non-approximability result is for the *single string* MLM problem.
4. The probabilistic automaton is often formulated as a probabilistic acceptor in the literature. Here we view PAs as generators. Thus the stochastic condition in this definition states that, for each state, the total probability of transitions out of that state sums to one, rather then the total probability for each state-letter pair as is the case for PAs as acceptors. Tzeng considers the incomparable problem of learning PAs as acceptors from queries (Tzeng, 1989).
5. In Physics, it is customary to use the natural logarithm for the definition of entropy. We use the binary logarithm for the entropy and the Kullback-Leibler divergence as is done in information and coding theory.
6. Note that the ideal code may have codeword lengths that are not integers.
7. We have also implicitly extended the notion of divergence for the generalized notion of probability distributions in which the total sum of probabilities over the domain may be less than one.
8. Note that these bounds can be made slightly larger, which would result in a slight improvement on the sample complexity. In particular, the bound on the transition probabilities can be $1/t^*m$ where t^* denotes the maximum number of transitions in G out of any state. The bound on the initial probabilities can be $1/|I|m$.
9. Thus, a letter group consists of either one or two letters.
10. This has recently been done in Abe, Takeuchi and Warmuth (1991) for various classes of probabilistic concepts with respect to both the Kullback-Leibler divergence and the quadratic distance.
11. Each path *for* the string u can be formalized as a length 6 sequence from $S \times S \times \Sigma_n$, where each ith member bas the ith symbol in u as its third component (its label) and consecutive transitions end and start in the same state. Here we simplify our notation by viewing a path simply as a sequence of states, leaving the labelings by u implicit.

References

Abe, N., Takeuchi, J., & Warmuth, M.K., (1991). Polynomial learnability of probabilistic concepts with respect to the Kullback-Leibler divergence. In *Proceedings of the 1991 Workshop on Computational Learning Theory.* San Mateo, CA: Morgan Kaufmann.

Angluin, D., (1978). On the complexity of minimal inference of regular sets. *Information and Control, 39,* 337–350.

Angluin, D., (1988). *Identifying languages from stochastic examples* (Technical Report YALEU/DCS/RR-614). Yale University.

Angluin, D., (1989). *Minimum consistent 2-state DFA problem is NP-complete.* Unpublished manuscript.

Barron, A.R. & Cover, T.M., (1989). Minimum complexity density estimation. *IEEE Transactions on Information Theory.*

Baum, L.E., (1972). An inequality and associated maximizaton technique in statistical estimation for probabilistic functions of a Markov process. *Inequalities, 3,* 1–8.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM, 36,* 929–965.

Gill, J., (1977) Probabilistic Turing machines. *SIAM J. Comput., 6,* 675–695.

Gold, E.M., (1978). Complexity of automaton identification from given data. *Information and Control, 37,* 302–320.

Hamming, R.W., (1986). Coding and Information Theory, Second Edition. Prentice-Hall.

Haussler, D., (1991). Decision theoretic generalizing of the pac model for neural net and other learning applications. *Information and Computation.* To appear. (An extended abstract appeared in the Proceedings of FOCS '89.)

Kearns, M., & Schapire, R., (1990). Efficient distribution-free learning of probabilitic concepts. In *Proceedings of IEEE Symposium on Foundations of Computer Science.*

Kullback, S., (1967). A lower bound for discrimination in terms of variation. *IEEE Transactions on Information Theory,* 126-127.

Laird, P.D. (1988). Efficient unsupervised learning. In *Proceedings of the 1988 Workshop on Computational Learning Theory.* San Mateo, CA: Morgan Kaufmann.

Levinson, S.E., Rabiner, L.R., & Sondhi, M.M., (1983). An introduction to the application of the theory of the probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal, 62.*

Pitt, L., & Warmuth, M.K., (1989). The minimum consistent DFA problem cannot be approximated within any polynomial. In *Proc. 19th ACM Symp. on Theory of Computation.* To appear in JACM.

Pollard, D., (1984). *Convergence of Stochastic Processes.* Springer-Verlag.

Tzeng, W., (1989). The equivalence and learning of probabilistic automata. In *Proceedings of the 30th IEEE Annual Symposium on the Foundations of Computer Science.*

Valiant, L.G., (1984). A theory of the learnable. *Communications of A.C.M., 27,* 1134-1142.

Yamanishi, K., (1991). A learning criterion for stochastic rules. *Machine Learning, 9,* .

A. Proofs of technical lemmas

In this appendix we prove the technical lemmas 5.2 through 5.5.

Proof of Lemma 5.2. Recall that $C[M](w) = \Pi_{H \in \Gamma_N}(M(*, *, H)/2)^{\#(H,w)}$ with the constraint that $\Sigma_{H \in \Gamma_n} M(*, *, H)/2 = 1$. By Lemma 4.2, the function of the form $f(x_1, \ldots, x_n) = x_1^{a_1} \cdot x_2^{a_2} \cdot \ldots \cdot x_n^{a_n}$ subject to the constraint $x_1 + x_2 + \ldots + x_n = c$ attains its maximum when $x_i = (a_i/\Sigma_{i=1}^n a_i) \cdot c$. Thus the maximum of $C[M](w)$ is obtained when $M(*, *, H)/2$ of each letter group H is set to the frequency of H in w. Thus $C[M](w) \leq \Pi_{H \in \Gamma_n}(\#(H, w)/|w|)^{\#(H,w)}$ and the product equals $C(w)$ by equality (5.6).

End of proof of Lemma 5.2.

Proof of Lemma 5.3. Let $M \in \mathcal{M}_n$ be an arbitrary stochastic matrix and $x = x_1 x_2 \ldots x_l \in \Sigma_n^*$ be an arbitrary string, and let l denote $|x|$, the length of x. For any k, $1 \leq k \leq l$, let $p_{k,j}$ denote the *conditional* probability that the machine M is in state j, *given* that it has just generated $x_1 x_2 \ldots x_{k-1}$. Note that $p_{k,0} + p_{k,1} = 1$ for any k. We can write $M(x_1 \ldots x_k)$ as follows:

$$M(x_1 \ldots x_k) \leq M(x_1 \ldots x_{k-1}) \cdot p_{k,0} \cdot M(0, *, x_k) + M(x_1 \ldots x_{k-1}) \cdot p_{k,1} \cdot M(1, *, x_k)$$

$$= M(x_1 \ldots x_{k-1})(p_{k,0} \cdot M(0, *, x_k) + p_{k,1} \cdot M(1, *, x_k))$$

$$\leq M(x_1 \ldots x_{k-1}) max\{M(0, *, x_k), M(1, *, x_k)\}, \text{ since } p_{k,0} + p_{k,1} = 1.$$

Hence it follows that:

$$M(x) \leq \prod_{i=1}^{|x|} max\{M(0, *, x_i), M(1, *, x_i)\} \tag{A.1}$$

Part 2 of the lemma follows from a similar argument. First, analogously to $p_{k,j}$, let $q_{k,j}$ denote the *conditional* probability that the machine M was in state j at time step k, *given* that it generated $x_{k+1}x_2 \ldots x_l$ after step k. Note that $q_{k,0} + q_{k,1} = 1$ for any k. We can write $M(x_k \ldots x_l)$ as follows:

$$\begin{aligned} M(x_k \ldots x_l) &\leq M(*, 0, x_k) \cdot q_{k,0} \cdot M(x_{k+1} \ldots x_l) + M(*, 1, x_k) \cdot q_{k,1} \cdot M(x_{k+1} \ldots x_l) \\ &= (q_{k,0} \cdot M(*, 0, x_k) + q_{k,1} \cdot M(*, 1, x_k))M(x_{k+1} \ldots x_l) \\ &\leq max\{M(*, 0, x_k), M(*, 1, x_k)\}M(x_{k+1} \ldots x_l), \text{ since } q_{k,0} + q_{k,1} = 1. \end{aligned}$$

Hence it follows that:

$$M(x) \leq \prod_{i=1}^{|x|} max\{M(*, 0, x_i), M(*, 1, x_i)\} \tag{A.2}$$

End of proof of Lemma 5.3.

Proof of Lemma 5.4. This lemma follows straightforwardly from Lemma 5.3. By the definition of $C[M](x)$,

$$C[M](x) = \prod_{H \in \Gamma_n} \left(\frac{M(*, *, H)}{2} \right)^{\#(H,x)}$$

But by Lemma 5.3, part 1, we have:

$$M(x) \leq \prod_{H \in \Gamma_n} (M(*, *, H))^{\#(H,x)}$$

So it follows:

$$M(x) \leq 2^{|x|} \cdot C[M](x)$$

End of proof of Lemma 5.4.

Proof of Lemma 5.5. Since $M_\tau(x) = 0$, in every path for x there must be at least one transition assigned zero probability by M_τ. Now since $\delta(M, M_\tau) = \Delta$, we have:

$$\lambda(M, M_\tau) = max_{H \in \Gamma_n, i \in S} \sum_{z \in H, j \in S, M_\tau(i,j,z)=0} \frac{M(i, j, z)}{M(i, *, H)} \leq \delta(M, M_\tau) = \Delta$$

Hence, in every path for x, there must be at least one transiton, say the k-th one, with probability not exceeding $\Delta \cdot max_{i \in S} M(i, *, H(x_k))$, where we wrote $H(x_k)$ for the letter group to which x_k belongs. But since $\nu(M, M_\tau) = max_{H \in \Gamma_n, i \in S} 2\,|M(i, *, H)/M(*, *, H) - 1/2\,| \leq \sqrt{\delta(M, M_\tau)} = \sqrt{\Delta}$,

$$\Delta \cdot max_{i \in S}\{M(i, *, H(x_k))\} \leq \Delta(1 + \sqrt{\Delta})\,\frac{M(*, *, H(x_k))}{2}.$$

Let $\Theta = S^{l+1}$ be the set of all paths of length $l = |x|$. Partition Θ into $\Theta_1, \Theta_2, \ldots, \Theta_l$ according to the first occurrence k of a transition probability not exceeding $\Delta(1 + \sqrt{\Delta})$ $M(*, *, H(x_k))/2$. Notice that $\Theta = \cup_{k=1}^{l} \Theta_k$ and Θ_k are mutually disjoint. Therefore if we define $M(x, \Theta_k) = \Sigma_{\langle i_0, \ldots, i_l \rangle \in \Theta_k} \Pi_{j=1}^{l} M(i_{j-1}, i_j, x_j)$, then we have $M(x) = \Sigma_{k=1}^{l} M(x, \Theta_k)$. Applying a similar argument as in the proof of Lemma 5.3, part 1, on each Θ_k, we obtain:

$$\begin{aligned} M(x, \Theta_k) &\leq \left(\prod_{j \neq k} max\{M(0, *, x_j), M(1, *, x_j)\}\right) \cdot \Delta(1 + \sqrt{\Delta})\,\frac{M(*, *, H(x_k))}{2} \\ &\leq \left(\prod_{j \neq k} (1 + \sqrt{\Delta})\,\frac{M(*, *, H(x_j))}{2}\right) \cdot \Delta(1 + \sqrt{\Delta})\,\frac{M(*, *, H(x_k))}{2} \\ &\leq \Delta(1 + \sqrt{\Delta})^l \cdot C[M](x) \end{aligned}$$

Hence we have:

$$M(x) \leq l\Delta(1 + \sqrt{\Delta})^l \cdot C[M](x) \tag{A.3}$$

End of proof of Lemma 5.5.

B. Proof of Lemma 5.6

In this appendix we prove Lemma 5.6, the key lemma used in the proof of Lemma 5.1, part 2. The proof uses the technical lemmas 5.2 through 5.5.

Proof of Lemma 5.6, part 1. Assume that $M(w) \geq 1/K\, C(w)$. For an arbitrary letter z, let i_z be the state $i \in S$ with the maximum out-share for z. (Let $i_z = 0$, if $M(0, *, z) = M(1, *, z)$.) Similary let j_z be the state with the maximum in-share for z. Then by Lemma 5.3, part 1, we have:

$$\begin{aligned} M(w_0) &= M((ab)^{k_0}) \\ &\leq M(i_a, *, a)^{k_0} \cdot M(i_b, *, b)^{k_0} \end{aligned}$$

$$\leq \left(\frac{M(q_a, *, a) + M(i_b, *, b)}{2}\right)^{2k_0}, \text{ because } xy \leq \left(\frac{x+y}{2}\right)^2 \ \forall x, y \in \mathbf{R} \tag{B.1}$$

Let Λ_{out} denote the fraction of the total (z, b)-share *not* given to the states with maximum out-shares for a and b, that is, $\Lambda_{out} = (\Sigma_{i \neq i_a} M(i, *, a) + \Sigma_{i \neq i_b} M(i, *, b))/M(*, *, (a, b))$. Alternatively, we can write Λ_{out} as $1 - (M(i_a, *, a) + M(i_b, *, b))/M(*, *, (a, b))$. Noting that $M(i_a, *, a) + M(i_b, *, b) = (1 - \Lambda_{out})M(*, *, (a, b))$, we obtain from (B.1):

$$M(w_0) \leq (1 - \Lambda_{out})^{2k_0}\left(\frac{M(*, *, (a, b))}{2}\right)^{2k_0}$$

$$= (1 - \Lambda_{out})^{2k_0} C[M](w_0), \text{ by definition of } C[M] \text{ and noting } w_0 = (ab)^{k_0}$$

Similarly, if we define Λ_{in} to be $\Sigma_{j \neq j_a} M(*, j, a) + \Sigma_{j \neq j_b} M(*, j, b)/M(*, *, (a, b)) = 1 - (M(*, j_a, a) + M(*, j_b, b))/M(*, *, (a, b))$, we obtain the following from Lemma 5.3, part 2.

$$M(w_0) \leq (1 - \Lambda_{in})^{2k_0} C[M](w_0) \tag{B.2}$$

So if we now let $\Lambda = max\{\Lambda_{out}, \Lambda_{in}\}$, then we have:

$$M(w_0) \leq (1 - \Lambda)^{2k_0} C[M](w_0) \tag{B.3}$$

Suppose for contradiction that $\Lambda > 2^{-11}/s$. Then by the choice of $k_0 = s2^{11}(\log K + |w_1 w_2 w_3 w_4 w_5|)$, we must have

$$M(w_0) < \left(1 - \frac{1}{s2^{11}}\right)^{2k_0} C[M](w_0) \tag{B.4}$$

$$< e^{-(2^{-11}/s)s2^{11}(\log K + |w_1 w_2 w_3 w_4 w_5|)} C[M](w_0)$$

$$< \frac{1}{K} \cdot \left(\frac{1}{2}\right)^{|w_1 w_2 w_3 w_4 w_5|} C[M](w_0) \tag{B.5}$$

Thus we can show via an argument of the style that was used in the proof of Lemma 5.1, part 2, given lemmas 5.2, 5.4 and 5.6:

$$M(w) \leq M(w_0)M(w_1 w_2 w_3 w_4 w_5)$$

$$< \frac{1}{K}\left(\frac{1}{2}\right)^{|w_1 w_2 w_3 w_4 w_5|} C[M](w_0) \cdot 2^{|w_1 w_2 w_3 w_4 w_5|} C[M](w_1 w_2 w_3 w_4 w_5), \text{ by Lemma 5.4}$$

$$\leq \frac{1}{K} C[M](w_0) C[M](w_1 w_2 w_3 w_4 w_5)$$

$$= \frac{1}{K} C[M](w)$$

$$\leq \frac{1}{K} C(w), \text{ by Lemma 5.2}$$

This contradicts our assumption, so we must conclude that $\Lambda \leq 2^{-11}/s$. Intuitively, we have shown that both the out-share and in-share of each of a and b must be highly concentrated on one of the two states, and hence that for each of a and b, there is exactly one transition in M which has almost all of the letter's share. Let $\sigma_a = (i_a, j_a)$ and $\sigma_b = (i_b, j_b)$ these two *dominating* transitions for a and b, respectively. Formally we can derive the following, recalling that $\Lambda_{out} = (\Sigma_{i \neq i_a} M(i, *, a) + \Sigma_{i \neq i_b} M(i, *, b))/M(*, *, (a, b))$, and $\Lambda_{in} = \Sigma_{j \neq j_a} (M(*, j, a) + \Sigma_{j \neq j_b} M(*, j, b))/M(*, *, (a, b))$

$$\begin{aligned}
\sum_{(i,j) \neq (i_a, j_a)} M(i, j, a) &\leq \sum_{i \neq i_a} M(i, *, a) + \sum_{j \neq j_a} M(*, j, a) \\
&\leq \Lambda_{out} M(*, *, (a, b)) + \Lambda_{in} M(*, *, (a, b)) \\
&\leq 2\Lambda M(*, *, (a, b)) \\
&\leq \frac{1}{s2^{10}} M(*, *, (a, b)) \qquad \text{(B.6)}
\end{aligned}$$

Similarly, for the letter b, we can show:

$$\sum_{(i,j) \neq (i_b, j_b)} M(i, j, b) \leq \frac{1}{s2^{10}} M(*, *, (a, b)) \qquad \text{(B.7)}$$

Next we will show that σ_a and σ_b must form a *cycle*, that is, $j_a = i_b$ and $j_b = i_a$. For suppose otherwise, then each of the 2^5 possible paths for the substring *abab* contains at least one transition other than σ_a and σ_b which by (B.6) and (B.7) has probability at most $2^{-10}/s$ $M(*, *, (a, b))$. Thus, each of the 2^5 paths has probability at most $2^{-10}/s$ $M(*, *, (a, b)) \cdot max\{M(*, *, a), M(*, *, b)\}^3$ and

$$\begin{aligned}
M(abab) &\leq 2^5 \cdot \frac{1}{s2^{10}} M(*, *, (a, b)) \cdot max\{M(*, *, a), M(*, *, b)\}^3 \\
&\leq \frac{1}{s2^5} M(*, *, (a, b))^4 \\
&\leq \frac{1}{2} \left(\frac{M(*, *, (a, b))}{2} \right)^4 = \frac{1}{2} C[M](abab) \qquad \text{(B.8)}
\end{aligned}$$

Recalling that $w_0 = (ab)^{k_0}$, the above implies that $M(w_0) \leq (1/2)^{k_0/2} C[M](w_0)$. As before, by the choice of k_0, this implies that $M(w) \leq 1/K\ C(w)$. We have thus shown that σ_a and σ_b must for a cycle. More precisely, we must have one of the following four possibilities.

1. $(\sigma_a, \sigma_b) = (0 \to 1, 1 \to 0)$.
2. $(\sigma_a, \sigma_b) = (0 \to 0, 0 \to 0)$.
3. $(\sigma_a, \sigma_b) = (1 \to 1, 1 \to 1)$.
4. $(\sigma_a, \sigma_b) = (1 \to 0, 0 \to 1)$.

Next, we will show that in fact only options 1 and 4 are possible, given the assumption that $M(w) \geq 1/K\ C(w)$: if we had either option 2 or 3, then we would have $M(w) < 1/K\ C(w)$. Assume without loss of generality that we have option 2, that is, $(\sigma_a, \sigma_b) = (0 \to 0, 0 \to 0)$. Then, note that $i_a = 0$ and $i_b = 0$. Hence we conclude:

$$M(i_a, *, a) + M(i_b, *, b) = M(0, *, a) + M(0, *, b) \leq 1 \tag{B.9}$$

Now let $\#(\overline{(a, b)}, w)$ be the number of all letters other than a and b in w and let h denote the inverse of the frequency of these letter in w, that is, $h = |w|/\#(\overline{(a, b)}, w)$. Note that $h > |w|/|w_1 w_2 w_3 w_4 w_5| > s2^{11}$ holds by the way w is defined. Also note that $M_\tau(*, *, (a, b))/2 = \#((a, b), w)/|w| = 1 - 1/h$. Using (B.9), we can derive:

$$\begin{aligned}
M(i_a, *, a)M(i_b, *, b) &\leq \left(\frac{M(i_a, *, a) + M(i_b, *, b)}{2}\right)^2, \text{ since } xy \leq \left(\frac{(x + y)}{2}\right)^2 \\
&\leq \left(\frac{1}{2}\right)^2, \text{ by (B.9)} \\
&= \left(\frac{1}{2} \frac{1}{1 - \frac{1}{h}} \frac{M_\tau(*, *, (a, b))}{2}\right)^2, \\
&\qquad \text{because } \frac{M_\tau(*, *, (a, b))}{2} = 1 - \frac{1}{h} \\
&= \frac{1}{4}\left(1 + \frac{1}{h - 1}\right)^2 \cdot \left(\frac{M_\tau(*, *, (a, b))}{2}\right)^2 \\
&\leq \frac{1}{2} \cdot \left(\frac{M_\tau(*, *, (a, b))}{2}\right)^2 \text{ noting that } h > s2^{11}.
\end{aligned} \tag{B.10}$$

Note that the above quantifies how much M loses on the letters a and b in w, as compared to a canonical matrix M_τ: M loses by at least a factor of 1/2 per each pair of a and b, if we have option 2. Next, we will bound how much M could possibly *gain* on the remaining letters as compared to a canonical matrix, by the fact that option 2 gives a and b *less* share. This will again be quantified in terms of h. From Lemma 5.3, part 1, we know that the

total generation probability assigned on these letters is at most $\Pi_{z\in\Sigma_n\setminus\{a,b\}} M(*, *, z)^{\#(z,w)}$ and this quantity is maximized when $M(*, *, z)$ are set proportionally to their frequencies in w within their total share $\Sigma_{z\in\Sigma_n\setminus\{a,b\}} M(*, *, z) = 2 - \Sigma_{z\in\{a,b\}} M(*, *, z) \leq 2$. Hence,

$$\prod_{z\in\Sigma_n\setminus\{a,b\}} M(*, *, z)^{\#(z,w)} \leq \prod_{z\in\Sigma_n\setminus\{a,b\}} \left(\frac{2\#(z, w)}{\#(\overline{(a, b)}, w)}\right)^{\#(z,w)} \tag{B.11}$$

But, by the definition of M_τ, we have that:

$$\prod_{z\in\Sigma_n\setminus\{a,b\}} M_\tau(*, *, z)^{\#(z,w)} = \prod_{z\in\Sigma_n\setminus\{a,b\}} \left(\frac{2\#(z, w)}{|w|}\right)^{\#(z,w)} \tag{B.12}$$

from (B.11) and (B.12), it follows that:

$$\begin{aligned}\prod_{z\in\Sigma_n\setminus\{a,b\}} M(*, *, z)^{\#(z,w)} &\leq \prod_{z\in\Sigma_n\setminus\{a,b\}} \left(\frac{|w|}{\#(\overline{(a, b)}, w)} \cdot \frac{2\#(z, w)}{|w|}\right)^{\#(z,w)} \\ &< \prod_{z\in\Sigma_n\setminus\{a,b\}} \left(\frac{|w|}{\#(\overline{(a, b)}, w)}\right)^{\#(z,w)} \\ &\quad \cdot \prod_{H\in\Gamma_n\setminus\{(a,b)\}} M_\tau(*, *, H)^{\#(H,w)} \\ &= (2h)^{\#(\overline{(a,b)},w)} \cdot \prod_{H\in\Gamma_n\setminus\{(a,b)\}} \left(\frac{M_\tau(*, *, H)}{2}\right)^{\#(H,w)}\end{aligned} \tag{B.13}$$

Here, the second to last inequality follows from the observation that by (5.4) for any partial letter z in letter group H, $M_\tau(*, *, H) = 4(\#(z, w)/|w|)$ and by (5.5) for any total letter z, $M_\tau(*, *, \{z\}) = 2(\#(z, w)/|w|)$. Now recalling that $C(w) = \Pi_{H\in\Gamma_n}(M_\tau(*, *, H)/2)^{\#(H,w)}$, we can use (B.10) and (B.13) to bound $M(w)$ from above by $1/K\ C(w)$, again contradicting our assumption.

$$M(w) \leq (M(i_a, *, a)M(i_b, *, b))^{\#(\{(a,b)\},w)/2} \prod_{z\in\Sigma_n\setminus\{a,b\}} M(*, *, z)^{\#(z,w)}, \text{ by Lemma 5.3, part 1}$$

$$\leq \left(\frac{1}{2}\right)^{\#((a,b),w)/2} \left(\frac{M_\tau(*, *, (a, b))}{2}\right)^{\#((a,b),w)} \cdot (2h)^{\#(\overline{(a,b)},w)}$$

$$\cdot \prod_{H\in\Gamma_n\setminus\{a,b\}} \left(\frac{M_\tau(*, *, H)}{2}\right)^{\#(H,w)} \text{ by (B.10) and (B.13)}$$

$$= \left(\frac{1}{2}\right)^{\#((a,b),w)/2} (2h)^{\#(\overline{(a,b)},w)} C(w), \text{ by definition of } C(w)$$

$$= \left(\left(\frac{1}{2} \right)^{(h-1)/2} 2h \right)^{\#(\overline{(a,b)},w)} C(w), \text{ since } \#((a, b), w) = (h - 1)\#(\overline{(a, b)}, w)$$

$$= \left(\left(\frac{1}{2} \right)^{(h-3-\log h)/2} \right)^{\#(\overline{(a,b)},w)} C(w)$$

$$\leq \left(\frac{1}{2} \right)^{\#(\overline{(a,b)},w)} C(w),$$

since $\left(\frac{1}{2} \right)^{h-3-\log h/2} < \frac{1}{2}$ because clearly $\frac{h - 3 - 2\log h}{2} > 1$ given $h > s2^{11}$

$$< \frac{1}{K} C(w), \text{ since } \#(\overline{(a, b)}, w) > \log K \tag{B.14}$$

Thus, we have verified that σ_a and σ_b are either $0 \to 1$ and $1 \to 0$, or $1 \to 0$ and $0 \to 1$. By (B.6) and (B.7), and the definition of $\lambda^{(a,b)}(M, \{a : 0 \to 1, b : 1 \to 0\})$ and $\lambda^{(a,b)}(M, \{a : 1 \to 0, b : 0 \to 1\})$, we conclude that one of the following must hold:

$$\lambda^{(a,b)}(M, \{a : 0 \to 1, b : 1 \to 0\}) \leq \frac{1}{s2^{10}}, \text{ or}$$

$$\lambda^{(a,b)}(M, \{a : 1 \to 0, b : 0 \to 1\}) \leq \frac{1}{s2^{10}}$$

Assume without loss of generality that the first of these two holds. Note that most of a's share is out of 0 and most of b's share is out of 1. Then, again by Lemma 5.3, part 1, we must have:

$$M(w_0) \leq (M(0, *, a)M(1, *, b))^{k_0}$$

$$\leq (1 - \nu^{(a,b)}(M)^2)^{k_0} \left(\frac{M(*, *, (a, b))}{2} \right)^{2k_0}, \text{ as shown in Example 5.1}$$

By an argument which is by now familiar, this implies that $\nu^{(a,b)}(M)^2 \leq 2^{-10}/s$. Hence, together with the earlier assumption that $\lambda^{(a,b)}(M, \{a : 0 \to 1, b : 1 \to 0\}) \leq 2^{-10}/s$, this implies that

$$\delta^{(a,b)}(M) = min_{\tau \in T_n} max\{\lambda^{(a,b)}(M, M_\tau), \nu^{(a,b)}(M)^2\}$$

$$\leq max\{\lambda^{(a,b)}(M, \{a : 0 \to 1, b : 1 \to 0\}), \nu^{(a,b)}(M)^2\}$$

$$\leq \frac{1}{s2^{10}}.$$

Proof of Lemma 5.6, part 2. This part follows directly from Lemma 5.3.

Proof of Lemma 5.6, part 3. This part follows at once if we establish the following claim, since $(2^{-10}/s) \leq 2^{-10}$ as long as the formula F is not empty ($s \geq 1$).

Claim B.1. *Let $u = afabfb$. Then, if $\delta^{(a,b)}(M) \leq (1/2)^{10}$ then $M(u) \leq (1 - \delta^{(a,b,f)}(M)/2)\, C[M](u)$.*

Proof of Claim B.1. What we wish to show is roughly as follows: Given that a and b go essentially as *intended*, that is a mostly goes from 0 to 1, and b mostly goes from 1 to 0, then f must also go essentially as intended, that is f must go mostly from 0 to 1, and 1 to 0. (In the mirror image of a canonical matrix, the roles of a and b are flipped, but here we assume without loss of generality that a goes from 0 to 1, and not the other way round.) We call these four transitions *intended transitions*, and all others *unintended transitions*. Note that $|u| = 6$, and so the length of each path (state sequence) for u is $|u| + 1 = 7$. Let Ω denote the set of all paths[11] possibly generating u, or $\Omega = S^7$. We then let Ω_1 denote the set of those paths in Ω containing only intended transitions (for u), except possibly at the two ends. For example, $1 \to 1 \to 0 \to 1 \to 0 \to 1 \to 1$ is in Ω_1 because the only unintended transitions in it are the first $1 \to 1$, labeled with a, and the last $1 \to 1$, labeled with b. On the other hand, $0 \to 1 \to 1 \to 1 \to 0 \to 1 \to 0$ is *not* in Ω_1 because the second transition, $1 \to 1$ labeled with f, and the third, $1 \to 1$ labeled with a are unintended. Let $\overline{\Omega_1}$ denote $\Omega \backslash \Omega_1$. Let $M(x, \Omega)$, in general, denote the probability that the string x is generated by one of the paths in Ω. That is, letting $x = x_1 \ldots x_l$ and $\omega = \langle i_0, \ldots, i_l \rangle$, $M(x, \Omega)$ is defined as follows.

$$M(x, \Omega) = \sum_{\langle i_0, \ldots, i_l \rangle \in \Omega} \prod_{j=1}^{l} M(i_{j-1}, i_j, x_j)$$

Then by the definition of $M(u)$ (Definition 2.6), Ω_1 and $\overline{\Omega_1}$ it follows that

$$M(u) \leq M(u, \Omega_1) + M(u, \overline{\Omega_1}) \tag{B.15}$$

Given that $\delta^{(a,b)}(M) \leq (1/2^{10})$, we will bound $M(u)$ from above as in the statement of the claim, by bounding from above the two terms $M(u, \Omega_1)$ and $M(u, \overline{\Omega_1})$ in (B.15) separately. For computing $M(u, \Omega_1)$, recall from Example 5.1 that:

$$M(0, 1, a)M(1, 0, b) \leq (1 - \delta^{(a,b)}(M)) \left(\frac{M(*, *, (a, b))}{2} \right)^2$$

Similarly we can also show

$$M(0, 1, f)M(1, 0, f) \leq (1 - \delta^{(f)}(M)) \left(\frac{M(*, *, f)}{2} \right)^2$$

Using these, we can bound $M(u, \Omega_1)$ as follows:

$$M(u, \Omega_1) \leq M(*, 1, a)M(1, 0, f)M(0, 1, a)M(1, 0, b)M(0, 1, f)M(1, *, b)$$

$$\leq M(*, *, a)M(*, *, b) \cdot M(0, 1, a)M(1, 0, b) \cdot M(0, 1, f)M(1, 0, f)$$

$$\leq \left(\frac{M(*, *, (a, b))}{2}\right)^2 \cdot (1 - \delta^{(a,b)}(M)) \left(\frac{M(*, *, (a, b))}{2}\right)^2 \cdot (1 - \delta^{(f)}(M)) \left(\frac{M(*, *, f)}{2}\right)^2$$

$$\leq (1 - \delta^{(f)}(M))(1 - \delta^{(a,b)}(M))C[M](u)$$

$$\leq (1 - \delta^{(a,bf)}(M))C[M](u) \tag{B.16}$$

The last inequality followed because $\delta^{(a,bf)}(M) = max\{\delta^{(f)}(M), \delta^{(a,b)}(M)\}$.

Now we bound from above the rest of $M(u)$, i.e., $M(u, \overline{\Omega_1})$. First note that by definition every path, say ω, in $\overline{\Omega_1}$ contains at least one unintended transition which is at neither end of u. The crucial observation is that because such an unintended transition has a transition before and after it, there must be at least *two consecutive unintended* transitions in ω. Now since $u = afabfb$, this implies that there must be at least one unintended (a, b)-transition, and another unintended transition in ω. Now note that there are at most five possible places for two consecutive unintended transitions in a path in $\overline{\Omega_1}$. If one fixes one of these five places as the place for the first occurrence of two consecutive unintended transitions, then the total probability of generating u via these paths is less than $C[M](u)$ by an approximate factor of $\lambda^{(a,b)}(M, M_\tau)$ times $\lambda^{(a,bf)}(M, M_\tau)$, accounting for the presence of unintended transitions at two specific positions, disregarding any favorable skews that might be present. For example, if we let $\overline{\Omega_{1,1}}$ denote the set of paths in which the first two consecutive transitions (for a and f) are unintended, then we can bound $M(u, \overline{\Omega_{1,1}})$ from above as follows. First, from a generalization of Lemma 5.3 to any subset of paths, we have:

$$M(u, \overline{\Omega_{1,1}}) \leq max\{M(0, 0, a), M(1, *, a)\} \cdot max_{i \in S} M(i, i, f) \cdot max_{i \in S} M(i, *, a) \cdot max_{i \in S} M(i, *, b) \cdot max_{i \in S} M(i, *, f) \cdot max_{i \in S} M(i, *, b) \tag{B.17}$$

Now by the definitions of leak and skew, if we let M_τ be an arbitrary canonical matrix, we have:

$$max\{M(0, 0, a), M(1, *, a)\} \leq max\{\lambda_0^{(a,b)}(M, M_\tau) \cdot M(0, *, (a, b)), \lambda_1^{(a,b)}(M, M_\tau) \cdot M(1, *, (a, b))\} \text{ by (5.9)}$$

$$\leq \lambda^{(a,b)}(M, M_\tau) \cdot max_{i \in S} M(i, *, (a, b)) \text{ by (5.10)}$$

$$\leq \lambda^{(a,b)}(M, M_\tau) \cdot (1 + \nu^{(a,b)}(M)) \cdot \frac{M(*, *, (a, b))}{2}$$

by (5.12)

Similarly,

$$max_{i \in S}\{M(i, i, f)\} \leq \lambda^{(f)}(M, M_\tau) \cdot (1 + \nu^{(f)}(M)) \cdot \frac{M(*, *, f)}{2}$$

For an arbitrary letter z in an arbitrary letter group H,

$$max_{i \in S} M(i, *, z) \leq (1 + \nu^{H}(M)) \cdot \frac{M(*, *, H)}{2} \text{ by (5.12)}$$

Plugging these into (B.17), we obtain:

$$M(u, \overline{\Omega_{1,1}}) \leq \lambda^{(a,b)}(M, M_\tau)(1 + \nu^{(a,b)}(M)) \frac{M(*, *, (a, b))}{2} \cdot$$

$$\lambda^{(f)}(M, M_\tau)(1 + \nu^{(f)}(M)) \frac{M(*, *, f)}{2} \cdot$$

$$(1 + \nu^{(a,b)}(M))^3 \left(\frac{M(*, *, (a, b))}{2}\right)^3 \cdot (1 + \nu^{(f)}(M)) \frac{M(*, *, f)}{2}$$

$$= \lambda^{(a,b)}(M, M_\tau) \cdot \lambda^{(f)}(M, M_\tau) \cdot (1 + \nu^{(a,b)}(M))^4 \cdot (1 + \nu^{(f)}(M))^2 \cdot$$

$$\left(\frac{M(*, *, (a, b))}{2}\right)^4 \cdot \left(\frac{M(*, *, f)}{2}\right)^2$$

$$\leq \lambda^{(a,b)}(M, M_\tau) \cdot \lambda^{(a,b,f)}(M, M_\tau) \cdot (1 + \nu^{(a,b)}(M))^4 \cdot (1 + \nu^{(f)}(M))^2 \cdot C[M](u)$$

Since we can derive the same inequality for each of the five possible places for the first two consecutive postions of unintended transitions, we obtain the following.

$$M(u, \overline{\Omega_1}) \leq 5\lambda^{(a,b)}(M, M_\tau)\lambda^{(a,b,f)}(M, M_\tau)(1 + \nu^{(a,b)}(M))^4(1 + \nu^{(f)}(M))^2 C[M](u)$$

Now since we have $\delta^{(a,b)}(M, M_\tau) = max\{\lambda^{(a,b)}(M, M_\tau), (\nu^{(a,b)}(M))^2\} \leq (1/2)^{10}$, and $1 + \nu^{(f)}(M) \leq 2$ by (5.13), we obtain:

$$M(u, \Omega_1) \leq 5 \cdot \left(\frac{1}{2}\right)^{10} \lambda^{(a,b,f)}(M, M_\tau) \left(1 + \left(\frac{1}{2}\right)^5\right)^4 2^2 C[M](u)$$

$$\leq \frac{1}{2}\lambda^{(a,b,f)}(M, M_\tau)C[M](u) \tag{B.18}$$

Putting (B.14), (B.16) and (B.18) together, we obtain:

$$\begin{aligned} M(u) &\leq M(u, \Omega_1) + M(u, \overline{\Omega_1}) \\ &\leq \left(1 - \delta^{(a,b,f)}(M) + \frac{1}{2}\lambda^{(a,b,f)}(M, M_\tau)\right) C[M](u) \\ &\leq \left(1 - \frac{\delta^{(a,b,f)}(M)}{2}\right) C[M](u) \end{aligned}$$

Proof of Lemma 5.6, part 4. This part follows from an analogous argument to the proof of Lemma 5.6, part 1, except the part that eliminates two of the four possible 'cycles,' since that part makes use of the fact that the frequency of a and b in w is very close to one. First, if we define Λ_i for each $i \leq n$, analogously to Λ in the earlier proof, then we can show that $max_{i \leq n}\Lambda_i \leq 2^{-10}/s$. We can also show that $(\nu^{(c_i,d_i)}(M))^2 \leq 2^{-10}/s$ by the same argument. So, it follows that if we let $\Delta = max_{i \leq n} min\{\delta^{(c_i,d_i)}(M, \{c_i : 0 \to 0, d_i : 0 \to 0\}), \delta^{(c_i,d_i)}(M, \{c_i : 0 \to 1, d_i : 1 \to 0\}), \delta^{(c_i,d_i)}(M, \{c_i : 1 \to 0, d_i : 0 \to 1\}), \delta^{(c_i,d_i)}(M, \{c_i : 1 \to 1, d_i : 1 \to 1\})\}\}$, then we have that if $M(w) \geq C(w)/K$ then $\Delta \leq 2^{-10}/s$.

Proof of Lemma 5.6, part 5. There are two cases. First, suppose that for some i, (c_i, d_i) pair is *not* either $(0 \to 1, 1 \to 0)$ or $(1 \to 0, 0 \to 1)$. More precisely, suppose that one of the following holds:

$$\Delta_i = \delta^{(c_i,d_i)}(M, \{c_i : 1 \to 1, d_i : 1 \to 1\}) \leq \frac{1}{s2^{10}} \tag{B.19}$$

$$\Delta_i = \delta^{(c_i,d_i)}(M, \{c_i : 0 \to 0, d_i : 0 \to 0\}) \leq \frac{1}{s2^{10}} \tag{B.20}$$

In this case, each of the 2^7 paths for the string $w_{3,i} = c_i f c_i d_i f d_i$ must contain at least two *unintended* transitions, that is, two transitions other than $f : 0 \to 1, f : 1 \to 0, c_i : 1 \to 1, d_i : 1 \to 1$ when (B.19) holds, and two transitions other than $f : 0 \to 1, f : 1 \to 0, c_i : 0 \to 0, d_i : 0 \to 0$, when (B.20) holds. Since all of Δ_i, $\lambda^{(f)}(M, M_\tau) \leq \delta^{(f)}(M, M_\tau)$, and $\nu^{(c,d,f)}(M) \leq max\{\Delta_i, \delta^{(f)}(M, M_\tau)\}$ are at most $2^{-10}/s$, we can bound $M(w_{3,i})$ from above as follows.

$$\begin{aligned} M(w_{3,i}) &\leq 2^7(max\{\Delta_i, \lambda^{(f)}(M, M_\tau)\})^2 \\ &\qquad \cdot (1 + \nu^{(c,d,f)}(M))^6 \left(\frac{M(*, *, (c_i, d_i))}{2}\right)^4 \left(\frac{M(*, *, f)}{2}\right)^2 \\ &\leq \frac{1}{2} C[M](w_{3,i}) \end{aligned}$$

It follows that in fact in this case $M(w_3) \leq (1/2)^{k_3} C[M](w_3)$, since each $w_{3,i}$ occurs k_3 times in w_3. Next, we assume that each (c_i, d_i) pair is set as intended, namely $(0 \to 1, 1 \to 0)$ or $(1 \to 0, 0 \to 1)$. Given this, we can prove the analogue of Claim B.1 in the proof of Lemma 5.6, part 3, where the string *afabfb* is replaced by $c_i f c_i d_i f d_i$. The desired conclusion follows at once.

Proof of Lemma 5.6, part 6. The proof of part 6 is similar to the proof of part 3. Using the same technique of dividing the set of paths for each substring of the form $u_i = a x_i b x_i c_i \bar{x}_i d_i \bar{x}_i beaeab$ we can obtain the following claim, exactly analogous to the claim in the earlier proof.

Claim B.2. *Let* $u_i = a x_i b x_i c_i \bar{x}_i d_i \bar{x}_i beaeab$. *Then, if* $\delta^{(a,b,c_i,d_i)}(M) \leq (1/2)^{10}$ *then* $M(u_i) \leq (1 - \delta^{(a,b,c_i,d_i,x_i,\bar{x}_i,e)}(M)/2) C[M](u_i)$.

If we divide the path set of u_i into Ω_1 and $\overline{\Omega_1}$, defined as before, then if we let M_τ be an arbitrary canonical matrix, we can show the following

$$
\begin{aligned}
M(u_i) &\leq M(u_i, \Omega_1) + M(u_i, \overline{\Omega_1}) \\
&\leq \left(1 - \delta^{(a,b,c_i,d_i,x_i,\bar{x}_i,e)}(M) + \frac{1}{2} \lambda^{(a,b,c_i,d_i,x_i,\bar{x}_i,e)}(M, M_\tau)\right) C[M](u) \\
&= \left(1 - \frac{\delta^{(a,b,c_i,d_i,x_i,\bar{x}_i,e)}(M))}{2}\right) C[M](u)
\end{aligned}
$$

Proof of Lemma 5.6, part 7. As we observed in the proof sketch of Theorem 5.1, if F is unsatisfiable, then for any canonical stochastic matrix M_τ, we must have $M_\tau(w_5) = 0$. But we have shown that $\delta(M) \leq 2^{-10}/s$. In other words, for some particular truth assignment τ, $\delta(M, M_\tau) \leq 2^{-10}/s$. Therfore we can apply Lemma 5.5 to each substring $u = \Pi_{j=1}^{s} ab(l_{j,1} l_{j,2} l_{j,3})b$ and obtain the following.

$$
M(u) \leq 6 \frac{1}{s 2^{10}} \left(1 + \sqrt{\frac{1}{s 2^{10}}}\right)^6 C[M](u)
$$

Clearly $M(u) \leq (1/2) C[M](u)$ holds, and thus we obtain $M(w_5) \leq (1/2)^{k_5} C[M](w_5)$.
End of proof of Lemma 5.6.