Noname manuscript No.
(will be inserted by the editor)

Statistical estimation with bounded memory

Leonid (Aryeh) Kontorovich

Received: date / Accepted: date

Abstract We investigate bounded-memory estimators of statistical functionals. It is shown that, for nondegenerate functionals and stochastic processes, it is impossible to achieve consistent estimation with bounded memory. In the positive direction, we show that $O(\log(1/\varepsilon))$ states suffice to achieve ε-consistent estimation for a natural class of functionals. A canonical optimal construction is conjectured for arbitrary statistical functionals.

Keywords statistical estimation · bounded memory · automaton · DFA · regular approximation

Mathematics Subject Classification (2000) 68Q70 · 62F12

1 Introduction

It is well known that the empirical average of independent, identically distributed (iid) random variables rapidly converges to their expectation. For concreteness, suppose that X_i, $i = 1, 2, \ldots$ are iid random variables with mean θ taking values in $[0, 1]$. Then it follows from Hoeffding's bound (Hoeffding, 1963) that for all $\varepsilon > 0$ we have

$$\mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i > \theta + \varepsilon\right\} \leq \exp(-2n\varepsilon^2), \tag{1}$$

Supported in part by the Israel Science Foundation

Aryeh Kontorovich
Department of Computer Science
Ben-Gurion University
Beer-Sheva 84105, ISRAEL
Tel.: +972-8-642-8048
Fax: +972-8-647-7650
E-mail: karyeh@cs.bgu.ac.il

with an analogous bound for the left tail. The inequality (1) guarantees the convergence of $A : (X_1, \ldots, X_n) \mapsto n^{-1}\sum X_i$ to θ in probability, implying that A is a consistent estimator of θ.

In this paper we address the feasibility of statistical estimation in bounded memory. The result, of course, will depend on the particular memory model we have in mind. A vast literature has dealt with analyzing the space complexity of various computations on data streams under different notions of memory (see the discussion in Section 3).

For the purpose of this paper, *memory* will be measured in the number of bits stored (or equivalently, in the number of tape squares accessed by a Turing machine). This choice of memory model is amply motivated in Cover (1969).

Our stringent definition precludes processing even a single infinite-precision real number, so there is no loss of generality in taking the X_i to be $\{0, 1\}$-valued. Assuming, as we do for now, that the X_i are iid, we have that X is a Bernoulli process, parametrized entirely by its mean $\theta = \mathbf{E}X_i = \mathbf{P}\{X_i = 1\} = 1 - \mathbf{P}\{X_i = 0\}$.

The empirical average in (1) may be computed naively by summing the n bits $X_1, \ldots, X_n$ and dividing by n. A standard trick circumvents storing the entire bit sequence, by initializing $A_1 := X_1$ and updating

$$A_{n+1} := \frac{nA_n + X_{n+1}}{n+1}. \tag{2}$$

This trick is infeasible in our model of memory since storing the integer n requires $\Omega(\log n)$ bits. There remains, however, the possibility that some other scheme performs consistent estimation in bounded memory. Thus, rather than analyzing the behavior of a particular function, such as the empirical mean, on the data stream, we ask whether *any* function computable in bounded

memory can be a consistent estimator of some distribution parameter.[1]

A *consistent estimator* of the Bernoulli parameter θ is a function $A : \{0,1\}^* \to \mathbb{R}$ such that $A(X_1, \ldots, X_n)$ converges in probability to θ:

$$\lim_{n\to\infty} \mathbf{P}\{|A(X, \ldots, X_n) - \theta| > \varepsilon\} = 0 \tag{3}$$

for all $\varepsilon > 0$. An obvious obstacle to achieving (3) is again the issue of *precision*: if θ takes values in some infinite set Θ, then clearly (3) is impossible for any function A computable with $O(1)$ bits of memory (as the latter can only distinguish among finitely many possibilities).

A central result of this paper is that, even when the precision obstacle is removed, statistical estimation with bounded memory remains impossible. In particular, we prove

Theorem 1 *Suppose that X_i, $i \in \mathbb{N}$ are Bernoulli random variables with parameter θ taking values in a fixed finite set Θ that contains distinct $\theta_0, \theta_1 \in (0,1)$. Then there is no consistent estimator for θ computable using $B = B(\Theta) < \infty$ bits of memory, where B is allowed to depend on Θ arbitrarily.*

This claim follows directly from a much more general result in Theorem 2 on the impossibility of estimating any nontrivial statistical functional of a Bernoulli process in bounded memory. In Theorem 3 this result is generalized further to a considerably broader class of random processes, including stationary ones with full support. We also investigate a partial converse to Theorem 1, where we construct ε-consistent estimators realized as DFAs with $O(\log(1/\varepsilon))$ states (Theorem 4).

The main contributions of this paper are Theorem 3 and the counterexamples when its conditions are not met. We also initiate here the study of canonical automata for estimating various statistical functionals.

2 Outline of paper

This paper is organized as follows. We review some relevant previous work and relate it to the present paper in Section 3 and set down the terminology used throughout the paper in Section 4. The main negative results are stated and proved in Section 5 for the Bernoulli case and generalized to all stationary processes with full support in Section 6. Counterexamples when the conditions of the main theorem fail to hold are given in Section 7 for some pathological random processes. In Section 8 we give some results on approximate statistical estimation with DFAs. Finally, we give a brief recap and outline a program of research for canonical finite-state estimators in Section 9.

[1] This question was posed to us by Ronen Brafman, motivated by problems in Reinforcement Learning (Brafman and Tennenholtz, 2002).

3 Background and related work

The problem of efficiently (in time and space) extracting useful information from a long sequence of data goes under the general heading of *streaming algorithms*. It appears that the earliest results along these lines were due to Morris (1978) and Flajolet and Martin (1985), who showed, roughly speaking, that logarithmic space suffices for approximating the total count and the number of unique items in a data stream. The field saw a surge of activity in the 1990's, including the seminal papers of Alon et al. (1999), Henzinger et al. (1999), and Feigenbaum et al. (2002), among many others. See Guha and McGregor (2009) for a recent result concerning quantile estimation and Muthukrishnan (2005) for a comprehensive survey of the subject.

An even earlier approach, beginning with Cover (1969), seeks to estimate various statistical quantities under assorted memory constraints. This line of work is continued in Hellman and Cover (1970); Hellman (1970); Hellman and Cover (1971); Cover et al. (1976). In this series of papers, the authors solve the hypothesis testing problem of the type $H_k : \theta = p_k$, and construct optimal deterministic and stochastic automata for this task. Following up, Lakshmanan and Chandrasekaran (1979) consider the hypothesis test $H_0 : \theta \geq p_0$ vs. $H_1 : \theta \leq p_1$, under the assumption that $p_0 > p_1$ and $\theta \notin (p_1, p_0)$. In the language of hypothesis testing, the present paper considers the feasibility of the test $H_0 : T(\theta) = 0$ vs. $H_1 : T(\theta) = 1$ for various predicates $T : \theta \mapsto \{0,1\}$ by means of deterministic automata. Closer in spirit to the present paper are Samaniego (1973) and Leighton and Rivest (1983), who deal directly with finite-memory parameter estimation.

A rather different line of research investigates approximations of non-regular languages by finite automata; see Eisman and Ravikumar (2005), Cordy and Salomaa (2007) and the references therein. Eisman and Ravikumar (2005) apparently made the first explicit connection between streaming algorithms and regular approximations. Indeed, since any bounded-memory algorithm can be implemented as a finite automaton, Theorem 1 may be recast as the following claim: if $A : \{0,1\}^* \to \Theta$ is a consistent estimator for the Bernoulli parameter $\theta \in \Theta$, then the language

$$A^{-1}(\{t\}) = \left\{x \in \{0,1\}^* : A(x) = t\right\}$$

is not regular for $0 < t < 1$.

Theorem 1 is also equivalent to the statement that no consistent estimator for $\mathbb{1}_{\{\theta>a\}}$, $a \in (0,1)$ can have a regular support set. For $a = \frac{1}{2}$, this last formulation is deceptively similar to a result of Eisman and Ravikumar (2005), quoted here in Theorem 7, which states, roughly, that no regular language can approximate the majority language on a set whose *unbiased* Bernoulli measure exceeds one-half. However, as we show in Section 8, this is not true for *biased* Bernoulli measures. Nevertheless, the main ingredient in our proofs — Markov chain analysis on the states of the DFA — was largely inspired by the technique of Eisman and Ravikumar (2005).

4 Notation

We follow the standard conventions for sets, languages, probability and automata. Thus, $\mathcal{X}$ is a finite alphabet, $\mathcal{X}^*$ is the set of all finite strings over $\mathcal{X}$, and a *language* is any $L \subseteq \mathcal{X}^*$. String length is denoted by $|\cdot|$ and

$$\mathcal{X}^{\leq k} = \cup_{i=0}^{k} \mathcal{X}^i = \{x \in \mathcal{X}^* : |x| \leq k\}.$$

For any $f : \mathcal{X}^* \to \{0,1\}$, its *support set* is

$$\mathrm{supp}(f) = f^{-1}(\{1\}) = \{x \in \mathcal{X}^* : f(x) = 1\}.$$

We use $\mathbb{N}$ to denote the natural numbers $\{1, 2, \ldots\}$.

A Deterministic Finite-state Automaton (DFA) over $\mathcal{X}$ is defined as the tuple $A = (Q, q_0, F, \delta)$ where

- $Q \subset \mathbb{N}$ is a finite set of states
- $q_0 \in Q$ is the starting state
- $F \subseteq Q$ is the set of the accepting states
- $\delta : Q \times \mathcal{X} \to Q$ is the deterministic transition function.

(Among the standard introductory texts on automata are Lewis and Papadimitriou (1981) and Sipser (2005).) The transition function δ may be extended to $Q \times \mathcal{X}^*$ via the recursion

$$\delta(q, (u_0, u_1, \ldots, u_n)) = \delta(\delta(q, (u_0, u_1, \ldots, u_{n-1})), u_n). \tag{4}$$

We regularly blur the distinction between an automaton and its underlying (multi)graph, whose vertex set is Q and whose edges are induced by δ. We also abuse the notation slightly by identifying languages $L \subseteq \mathcal{X}^*$ and automata A with their characteristic functions $L, A : \mathcal{X}^* \to \{0,1\}$, defined by $L(x) = \mathbb{1}_{\{x \in L\}}$ (resp., $A(x) = \mathbb{1}_{\{\delta(q_0,x) \in F\}}$). The probability $\mathbf{P}\{\cdot\}$ is always defined with respect to the random process X specified in context, and we often use the shorthand $X^n \equiv (X_1, X_2, \ldots, X_n)$. The Bernoulli process X has *parameter* θ if the X_i are iid with

$$\mathbf{P}\{X_i = 1\} = \theta = 1 - \mathbf{P}\{X_i = 0\}.$$

5 The Bernoulli case

We begin with the problem of estimating the Bernoulli parameter θ. Although the result in this section is subsumed by the more general Theorem 3, we present Theorem 2 first for expositional clarity.

In this section, the alphabet is $\mathcal{X} = \{0,1\}$ and a *statistical predicate* $T : [0,1] \to \{0,1\}$ is any binary map acting on the Bernoulli parameter (e.g., $T(\theta) = \mathbb{1}_{\{\theta>1/2\}}$). T is *nontrivial* if it is not identically 0 or 1. As before, a *consistent* estimator for T is any $A : \{0,1\}^* \to \{0,1\}$ such that $A(X^n)$ converges in probability to $T(\theta)$:

$$\lim_{n\to\infty} \mathbf{P}\{A(X^n) \neq T(\theta)\} = 0. \tag{5}$$

We show that finite automata cannot yield consistent estimators, which in particular implies Theorem 1:

Theorem 2 *Suppose X is a Bernoulli process with parameter $\theta \in (0,1)$ and T is a nontrivial statistical predicate. If A is a consistent estimator for T, then its support language,* $\mathrm{supp}(A)$*, is not regular.*

Remark: This theorem was proved together with Gerald Eisman.

Proof Suppose to the contrary that the language $\mathrm{supp}(A)$ is regular. Then it is recognized by some DFA $A = (Q, \delta, q_0, F)$.

The Bernoulli process X together with the transition function δ define a Markovian dynamics on Q as follows. Defining the random variable $\xi_n \in Q$ to be the state of the automaton after reading the random string $(X_1, \ldots, X_n)$, we have $\mathbf{P}\{\xi_0 = q_0\} = 1$ and

$$\mathbf{P}\{\xi_{n+1} = q' \,|\, \xi_n = q\} = \begin{cases} \theta, & \delta(q,1) = q' \\ 1-\theta, & \delta(q,0) = q' \\ 0, & \text{otherwise} \end{cases}$$

(in the degenerate case that $\delta(q,0) = \delta(q,1) = q'$, $\mathbf{P}\{\xi_{n+1} = q' \,|\, \xi_n = q\} = 1$).

Since T is nontrivial, there are $\theta_0, \theta_1 \in (0,1)$ such that $T(\theta_0) = 0$ and $T(\theta_1) = 1$. Thus, consistency as defined in (5) implies

$$\lim_{n\to\infty} \mathbf{P}\{\xi_n \in F\} = T(\theta), \qquad \theta \in \{\theta_0, \theta_1\}, \tag{6}$$

which says that the probability that the DFA A is in an accepting state approaches either 0 or 1, depending on whether $\theta = \theta_0$ or $\theta = \theta_1$.

The Markov chain $\{\xi_n\}$ has a finite state space Q, which decomposes as $Q = E \cup H$, $E \cap H = \emptyset$ where E are the ergodic states and H are the transient states (see Kemeny and Snell (1976) for general facts about

finite Markov chains). The crucial observation is that E and H are determined by the automaton A alone, and do not depend on θ. Indeed, θ is assumed different from 0 and 1, and for all $\theta \in (0,1)$, the directed graphs underlying the Markov chain $\{\xi_n\}$ are identical. A state q is transient if the Markov chain will almost surely visit q only finitely many times — and so transience does not depend on the actual value of θ. Thus, the pair (E,H) is the same for $\theta = \theta_0$ and $\theta = \theta_1$. Now by definition of transience and ergodicity

$$\lim_{n\to\infty} \mathbf{P}\{\xi_n \in H\} = 0$$

while

$$\lim_{n\to\infty} \mathbf{P}\{\xi_n = q\} > 0$$

for all $q \in E$. Thus $\lim_{n\to\infty} \mathbf{P}\{\xi_n \in F\} = 0$ if and only if $F \subseteq H$. Conversely, $\lim_{n\to\infty} \mathbf{P}\{\xi_n \in F\} = 1$ if and only if $E \subseteq F$. These properties hold simultaneously for $\theta = \theta_0$ and $\theta = \theta_1$, and thus a DFA distinguishing between these two scenarios, as in (6), is impossible. □

6 Stationary processes with full support

We refer the reader to Kallenberg (2002) for the relevant background on random processes. Let $\mathcal{X}$ be a finite set and let $\Omega = \mathcal{X}^{\mathbb{N}}$ be the set of all $\mathcal{X}$-valued infinite sequences. We equip Ω with the usual σ-algebra $\mathscr{F}$ induced by the finite-dimensional cylinders; in everything that follows, all measures on Ω are defined on the measurable space $(\Omega, \mathscr{F})$. A probability measure μ on Ω induces the $\mathcal{X}$-valued coordinate process $X = (X_1, X_2, \ldots)$ via the identity

$$\mathbf{P}\{X \in A\} = \mu(A) \tag{7}$$

for all $A \in \mathscr{F}$.

The process $X = (X_1, X_2, \ldots)$ is said to be *stationary* if

$$\mathbf{P}\{(X_{t_1}, X_{t_2}, \ldots, X_{t_m}) = x\} = \mathbf{P}\{(X_{t_1+k}, X_{t_2+k}, \ldots, X_{t_m+k}) = x\}$$

for all $k, m \geq 1$, all $0 < t_1 < \ldots < t_k$, and all $x \in \mathcal{X}^m$. Furthermore, X has *full support* if every realization occurs with positive probability:

$$\mathbf{P}\{(X_1, \ldots, X_n) = x\} \equiv \mu(\{x\}) > 0 \tag{8}$$

for all $n \geq 1$ and all $x \in \mathcal{X}^n$.

The extension of Theorem 2 to the much broader class of stationary processes with full support requires a bit of additional abstraction. Let $\mathcal{M}$ be some collection of probability measures on Ω, where each $\mu \in \mathcal{M}$ is uniquely identified with the $\mathcal{X}$-valued random process $X = (X_1, X_2, \ldots)$ as in (7). A *statistical predicate* is any mapping $T : \mathcal{M} \to \{0,1\}$, and T is *nontrivial* if there are $\mu_0, \mu_1 \in \mathcal{M}$ such that $T(\mu_0) \neq T(\mu_1)$. Finally, $A : \mathcal{X}^* \to \{0,1\}$ is a *consistent* estimator for T if

$$\lim_{n\to\infty} \mu\{x \in \mathcal{X}^n : A(x) \neq T(\mu)\} = 0 \tag{9}$$

for all $\mu \in \mathcal{M}$; a shorthand way of writing the above is $\mu\{A(X^n) \neq T(\mu)\} \to 0$.

We are ready to proceed with the generalization:

Theorem 3 *Let $\mathcal{M}$ be a collection of stationary $\mathcal{X}$-valued processes with full support and suppose $T : \mathcal{M} \to \{0,1\}$ is a nontrivial predicate. Then a consistent estimator for T cannot have regular support.*

Proof Assume to the contrary that T has an estimator with a regular support set, and that the latter is recognized by some DFA $A = (Q, \delta, q_0, F)$. We may take A to be a minimal DFA, and in particular, every state is reachable from the starting state q_0:

$$\forall q \in Q \ \exists u \in \mathcal{X}^* : q = \delta(q_0, u) \tag{10}$$

(δ is defined recursively in (4)). The $\mathcal{X}$-valued process $(X_1, X_2, \ldots)$ induces the Q-valued process $\xi = \xi(X)$ defined formally by $\xi_0 = q_0$ and

$$\xi_n = \delta(q_0, (X_1, \ldots, X_n))$$

for $n \geq 1$. In words, $\xi = (\xi_0, \xi_1, \ldots)$ is the sequence of states visited by the automaton when reading X. An immediate consequence of (10) and the full-support assumption (8) is that for all $\mu \in \mathcal{M}$, any state $q \in Q$ is reachable with positive probability:

$$\forall \mu \in \mathcal{M} \ \forall q \in Q \ \exists n \in \mathbb{N} : \mu\{\xi_n(X) = q\} > 0. \tag{11}$$

Since T is nontrivial, there are measures $\mu_0, \mu_1 \in \mathcal{M}$ such that $T(\mu_0) = 0$ and $T(\mu_1) = 1$. Thus the consistency property in (9) implies

$$\lim_{n\to\infty} \mu_j\{\xi_n(X) \in F\} = j \tag{12}$$

for $j \in \{0,1\}$. An obvious consequence of (12) is that $\emptyset \neq F \neq Q$.

A *directed path* from $q \in Q$ to $q' \in Q$ is a sequence $\pi \in \mathcal{X}^*$ such that $\delta(q, \pi) = q'$. We say that π is a minimal directed path from q to q' if there is no directed path from q to q' shorter than π. For $E, E' \subset Q$, a directed path from E to E' is a directed path from any $q \in E$ to any $q' \in E'$.

We claim that for every $f \in F$ there is a directed path to some $g \in Q \setminus F$, and vice versa. Suppose to the contrary that there is some $f \in F$ with no directed

paths to any member of $Q \setminus F$. By (11), this f is reachable from q_0 with positive μ_0-probability. Thus, if there were no directed path from f to any $Q \setminus F$, the condition $\lim_{n\to\infty} \mu_0 \{\xi_n(X) \in F\} = 0$ would be violated. An analogous argument shows that there must be a directed path from every $g \in Q \setminus F$ to some $f \in F$.

Let Π be the (finite) collection of all minimal directed paths from F to $Q \setminus F$. Stationarity and full-support imply that

$$\mu\left\{(X_t, X_{t+1}, \ldots, X_{t+|\pi|-1}) = \pi\right\} > 0 \tag{13}$$

for all $t \in \mathbb{N}$, $\mu \in \mathcal{M}$ and all $\pi \in \Pi$; furthermore, this quantity depends only on μ and π — but not on t. Thus $\mu_1(\{\pi\})$ is well-defined, as is $\varepsilon_1 = \min_{\pi\in\Pi} \mu_1(\{\pi\}) > 0$.

We claim that

$$\liminf_{n\to\infty} \mu_1 \{\xi_n(X) \in F\} \leq 1 - \varepsilon_1. \tag{14}$$

This holds because from every $f \in F$ there is a minimal path leading to some $g \in Q \setminus F$, which has μ_1-probability at least ε_1. But (14) violates the consistency condition $\lim_{n\to\infty} \mu_1 \{\xi_n(X) \in F\} = 1$. It follows that no DFA can distinguish μ_0 from μ_1 with probability approaching 1. □

7 Counterexamples

Theorem 3 gives sufficient conditions for a random process not to admit consistent finite-state statistical estimators. In this section, we give examples of such estimators for processes violating the conditions of stationarity and full-support. The basic intuition is that a DFA cannot accumulate statistical information — it can only be driven into a certain state by the process (as in Section 7.1), or exploit forbidden patterns (as in Section 7.2).

7.1 Non-stationary process

For $j \in \{0, 1\}$, let $\mathcal{M}_j$ be the collection of $\{0, 1\}$-valued processes $X = (X_1, X_2, \ldots)$ where each realization satisfies

$$\lim_{n\to\infty} \mathbf{P}\{X_n = j\} = 1$$

and define $\mathcal{M} = \mathcal{M}_0 \cup \mathcal{M}_1$. Thus, every process in $\mathcal{M}$ eventually becomes dominated entirely by 0s or 1s. Consider the statistical predicate $T : \mathcal{M} \to \{0, 1\}$ that distinguishes the 0-dominated processes from the 1-dominated ones:

$$T(\mu) = \mathbb{1}_{\{\mu\in\mathcal{M}_1\}}.$$

The processes in $\mathcal{M}$ are clearly not stationary, and a simple automaton (Figure 1) realizes the predicate T. It

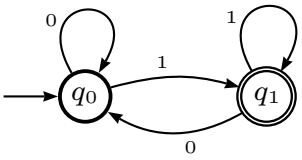

Fig. 1 A two-state automaton distinguishes 0-dominant processes from 1-dominant ones.

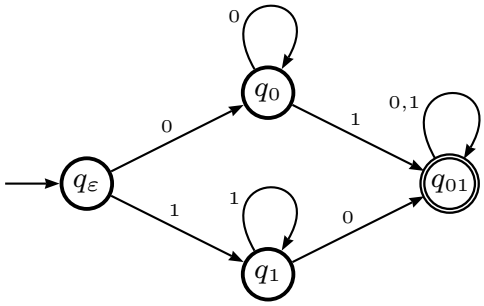

Fig. 2 A four-state automaton distinguishes degenerate processes from nondegenerate ones.

is not difficult to verify that this automaton will occupy state q_j with probability approaching 1 when reading the realization of a j-dominant process, for $j \in \{0, 1\}$.

7.2 Process without full support

Let $\mathcal{M}$ be the collection of $\{0, 1\}$-valued iid Bernoulli processes $X = (X_1, X_2, \ldots)$ with parameter $\theta \in [0, 1]$. Let us call the processes with $\theta \in \{0, 1\}$ *degenerate* and those with $\theta \in (0, 1)$ *nondegenerate* and define the predicate $T : \mathcal{M} \to \{0, 1\}$ to be 1 if μ is nondegenerate and 0 otherwise. The processes comprising $\mathcal{M}$ are stationary but do not all have full support, and a simple automaton (Figure 2) realizes the predicate T. It is easily verified that this automaton will occupy state q_{01} with probability approaching 1 when reading a nondegenerate process and will become trapped either in state q_0 or q_1 when reading a degenerate process.

8 Approximate statistics with a DFA

We revisit the problem of approximating the Bernoulli parameter θ with a DFA. As discussed in Section 1, this question is only meaningful if θ is allowed to take values in some finite set Θ. Suppose for concreteness that $\Theta = \{0 < \theta_0 < \theta_1 < \ldots < \theta_k < 1\}$. Then the problem of determining whether $\theta = \theta_j \in \Theta$ is reduced to deciding whether $\theta > \theta_{j-1}$ and $\theta < \theta_{j+1}$.[2] Of course, by Theorem 2, a consistent estimator for $T_a(\theta) = \mathbb{1}_{\{\theta>a\}}$ cannot

[2] Leighton and Rivest (1983) estimate θ directly, bypassing statistical predicates. Rajwan and Feder (2000) give an optimal quantization of the parameter space.

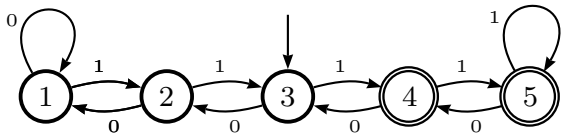

Fig. 3 The automaton $M(5)$ agrees with MAJ on all $x \in \{0,1\}^{\leq 4}$. The general $M(k) = (Q, q_0, F, \delta)$ is constructed as follows: $Q = \{1, \ldots, k\}$, $q_0 = \left\lfloor \frac{k+1}{2} \right\rfloor$, $F = \{\lceil k/2 \rceil + 1, \ldots, k\}$ and $\delta(i,0) = \max\{i-1, 1\}$, $\delta(i,1) = \min\{i+1, k\}$ for $1 \leq i \leq k$.

be realized by any DFA. Consider, however, relaxing the requirement of consistency in (5) to ε-consistency:

$$\limsup_{n\to\infty} \mathbf{P}\{A(X^n) \neq T(\theta)\} < \varepsilon. \tag{15}$$

We shall examine the case of $T_{1/2}$ in some detail. To this end, recall the majority function $\mathrm{MAJ} : \{0,1\}^* \to \{0,1\}$, defined by

$$\mathrm{MAJ}(x) = \mathbb{1}_{\left\{\sum_{i=1}^{|x|} x_i > \frac{1}{2}|x|\right\}}.$$

We observe that any consistent estimator of $\mathbb{1}_{\{\theta>1/2\}}$ must asymptotically agree with MAJ:

Theorem 4 *Let X be a Bernoulli process with parameter $\theta \neq \frac{1}{2}$ and suppose that $A : \{0,1\}^* \to \{0,1\}$ is a consistent estimator of the predicate $T_{1/2}(\theta) = \mathbb{1}_{\{\theta>1/2\}}$. Then*

$$\lim_{n\to\infty} \mathbf{P}\{A(X^n) \neq \mathrm{MAJ}(X^n)\} = 0.$$

Proof Assume without loss of generality that $\theta > \frac{1}{2}$, so $T_{1/2}(\theta) = 1$. Then,

$$\begin{aligned}
\mathbf{P}\{\mathrm{MAJ}(X^n) \neq 1\} &= \mathbf{P}\left\{\sum_{i=1}^n X_i \leq \frac{n}{2}\right\} \\
&= \mathbf{P}\left\{\frac{1}{n}\sum X_i \leq \theta - \left(\theta - \tfrac{1}{2}\right)\right\} \\
&\leq \exp(-2n(\theta - \tfrac{1}{2})^2), \qquad (16)
\end{aligned}$$

where the last inequality is Hoeffding's; a similar analysis holds for $\theta < \frac{1}{2}$. Thus,

$$\begin{aligned}
&\mathbf{P}\{A(X^n) \neq \mathrm{MAJ}(X^n)\} \leq \\
&\mathbf{P}\big\{A(X^n) \neq T_{1/2}(\theta)\big\} + \mathbf{P}\big\{\mathrm{MAJ}(X^n) \neq T_{1/2}(\theta)\big\} \to 0
\end{aligned}$$

where the first term goes to 0 because A is a consistent estimator for $T_{1/2}$ and the second term vanishes by (16). □

Let $M(k)$ be a minimal DFA[3] which agrees with MAJ on all binary strings of length less than k; these are illustrated in Figure 3. One might inquire how well $M(k)$ approximates $\mathbb{1}_{\{\theta>1/2\}}$ on long Bernoulli sequences, and the following theorem provides an answer:

Theorem 5 *Let X be a Bernoulli process with parameter θ let $M(k)$ be defined as above. Then*

$$\begin{aligned}
\varepsilon &\equiv \lim_{n\to\infty} \mathbf{P}\{X^n \notin M(k)\} \\
&= \frac{\rho^{\lfloor k/2 \rfloor + 1} - \rho}{\rho^{k+1} - \rho}, \qquad (17)
\end{aligned}$$

where $\rho = \theta/(1-\theta)$. For even k we have

$$\varepsilon \leq \frac{1}{2}(2 - 2\theta)^{k/2} \tag{18}$$

(which holds for all $\theta \in (0,1)$ but is vacuous for $\theta < 1/2$).

Remark: we thank Daniel Dadush and Daniel Berend for help with this calculation.

Proof The Bernoulli process X induces the Markov chain $\xi = (\xi_1, \ldots)$, $\xi_i \in Q$, on the states of $M(k)$ as described in the proof of Theorem 2. By construction of the DFA $M(k)$, the induced Markov chain is ergodic (for a visual illustration, relabel every "1" edge in Figure 3 with θ and every "0" edge with $1-\theta$). Its unique stationary distribution $\pi \in \mathbb{R}^k$ has the interpretation

$$\pi_q = \lim_{n\to\infty} \mathbf{P}\{\xi_n = q\}$$

and obeys the recurrence

$$\pi_i = \theta\pi_{\max\{i-1,1\}} + (1-\theta)\pi_{\min\{i+1,k\}}, \qquad 1 \leq i \leq k.$$

This relation is satisfied by the vector

$$\tilde{\pi}_i = \theta^i(1-\theta)^{k-i}, \qquad 1 \leq i \leq k,$$

which must be normalized to make it into a probability distribution. The accepting states $F \subset Q = \{1, \ldots, k\}$ of $M(k)$ are all $q > \lceil k/2 \rceil$, and so the limiting probability of being in a rejecting state is given by

$$\begin{aligned}
\sum_{q\notin F} \pi_q &= \frac{\sum_{i=1}^{\lfloor k/2 \rfloor + 1} \theta^i(1-\theta)^{k-i}}{\sum_{i=1}^{k} \theta^i(1-\theta)^{k-i}} \\
&= \frac{(1-\theta)^k \sum_{i=1}^{\lfloor k/2 \rfloor + 1} \left(\frac{\theta}{1-\theta}\right)^i}{(1-\theta)^k \sum_{i=1}^{k} \left(\frac{\theta}{1-\theta}\right)^i}
\end{aligned}$$

and the latter sum up as geometric series to yield (17). To obtain (18) from (17), set $k = 2\ell$. Then

$$\begin{aligned}
\frac{\rho^{\lfloor k/2 \rfloor + 1} - \rho}{\rho^{k+1} - \rho} &= \frac{\rho^{\ell+1} - \rho}{\rho^{2\ell+1} - \rho} = \frac{\rho^\ell - 1}{\rho^{2\ell} - 1} = \frac{\rho^\ell - 1}{(\rho^\ell + 1)(\rho^\ell - 1)} \\
&= \frac{1}{\rho^\ell + 1} = \frac{1}{(\theta/(1-\theta))^\ell + 1} \\
&= \frac{(1-\theta)^\ell}{\theta^\ell + (1-\theta)^\ell}.
\end{aligned}$$

[3] One may take the family of automata constructed in Figure 3 as the *definition* of $M(k)$ and prove as a simple exercise that there is no DFA with fewer states agreeing with MAJ on all of $\{0,1\}^{<k}$. For small values of k, the techniques of Trakhtenbrot and Barzdin' (1973) or Angluin (1987) may be used to construct a minimal DFA agreeing with a given membership oracle on $\{0,1\}^{<k}$.

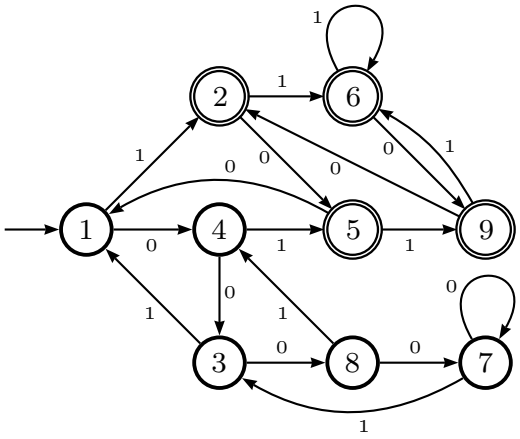

Fig. 4 The automaton $M_{1/3}(7)$ for approximating $T_{1/3}$.

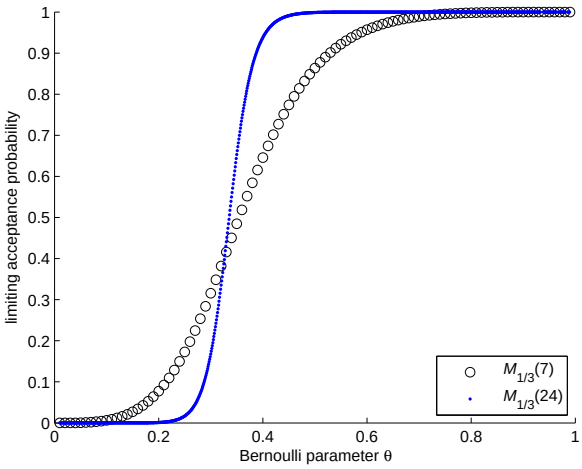

Fig. 5 The limiting behavior of $M_{1/3}(7)$ and $M_{1/3}(24)$, whose sizes are 9 and 29, respectively. The automata were obtained via Angluin's algorithm (Angluin, 1987).

Since the r.h.s. of (18) is $2^{\ell-1}(1-\theta)^\ell$, it remains to show that $\left(\theta^\ell + (1-\theta)^\ell\right)^{-1} \leq 2^{\ell-1}$. But this follows from Jensen's inequality:

$$\frac{\theta^\ell + (1-\theta)^\ell}{2} \geq \left(\frac{\theta}{2} + \frac{1-\theta}{2}\right)^\ell = 2^{-\ell}.$$

□

It follows from Theorem 5 that the DFA $M(2k)$ will disagree with the majority function on long runs of Bernoulli processes with parameter θ with probability at most

$$R(\theta, 2k) = \frac{1}{2}\left(1 - 2\left|\theta - \frac{1}{2}\right|\right)^k.$$

Note that

$$\lim_{k\to\infty} R(\theta, k) = 0, \qquad \theta \neq \frac{1}{2} \tag{19}$$

exponentially fast, while

$$\lim_{\theta\to 1/2} R(\theta, k) = \frac{1}{2}. \tag{20}$$

One additional issue of potential interest is the *mixing rate* of the Markov chain $\{\xi_i\}$ induced by the Bernoulli process X on the states of $M(k)$.

Theorem 6 *Let $\{\xi_i\}$ be the Markov chain defined in Theorem 5. Denote the marginal distribution of ξ_i by P_i and the stationary distribution by P_∞. Then, for even k, we have*

$$\left\|P_{nk/2} - P_\infty\right\|_{\text{TV}} \leq \exp(-n \min\{\theta, 1-\theta\}^{k/2}),$$

where $\|\cdot\|_{\text{TV}} = \frac{1}{2}\|\cdot\|_1$ is the total variation norm.

Remark: Note that, in contrast with (19) and (20), the convergence to the stationary distribution is most rapid for $\theta = 1/2$.

Proof Let A be the $(k+1)\times(k+1)$ column-stochastic matrix representing the transition kernel of the Markov chain in question; then A is given by $A_{1,1} = A_{i-1,i} = 1-\theta$ for $i \in [2, k+1]$ and $A_{k+1,k+1} = A_{i+1,i} = \theta$ for $i \in [1, k]$; all other A_{ij} are zero. For any column-stochastic matrix B, define $\tau(B)$ to be the maximum total-variation distance between any two columns of B; this is the Doeblin *contraction coefficient* of B. We claim that

$$\tau(A^{k/2}) = 1 - \min\{\theta, 1-\theta\}^{k/2}.$$

This is verified by recalling the interpretation of $[A^n]_{ij}$ as the probability of reaching state i from j in n steps. Since state $k/2+1$ is the only one that is reachable from both states 1 and $k+1$ in $k/2$ steps, with probabilities $\theta^{k/2}$ and $(1-\theta)^{k/2}$, respectively, we have that

$$\begin{aligned}\tau(A^{k/2}) &\geq \max\left\{1-\theta^{k/2}, 1-(1-\theta)^{k/2}\right\} \\ &= 1 - \min\{\theta, 1-\theta\}^{k/2};\end{aligned} \tag{21}$$

to see this, recall Scheffé's theorem, a special case of which states that for any two probability measures P, P' on the finite set Q, we have

$$\|P - P'\|_{\text{TV}} = \max_{E\subseteq Q} \sum_{q\in E} (P(q) - P'(q)).$$

It is not difficult to see that since states 1 and $k+1$ are the farthest from each other in terms of steps, the pairwise total variation distance is maximized by their respective columns, and thus the inequality in (21) is actually an equality. Finally, let us recall how the contraction coefficient got its name:

$$\|B(P - P')\|_{\text{TV}} \leq \tau(B)\,\|P - P'\|_{\text{TV}}$$

holds for any column-stochastic matrix B and two probability measures P, P' (see, for example, Bremaud (1999) or Kontorovich and Ramanan (2008)). Thus,

$$\begin{aligned}\left\|P_{nk/2} - P_\infty\right\|_{\text{TV}} &= \left\|A^{nk/2}P_0 - P_\infty\right\|_{\text{TV}} \\ &= \left\|A^{nk/2}(P_0 - P_\infty)\right\|_{\text{TV}} \\ &\le \tau(A^{k/2})^n \left\|P_0 - P_\infty\right\|_{\text{TV}} \\ &\le \tau(A^{k/2})^n \\ &= \left(1 - \min\left\{\theta, 1-\theta\right\}^{k/2}\right)^n \\ &\le \exp(-n \min\left\{\theta, 1-\theta\right\}^{k/2}),\end{aligned}$$

which completes the proof. □

The approach in Theorem 5 can be generalized to obtain approximate finite-state estimators for the Bernoulli statistical predicate $T_a(\theta) = \mathbb{1}_{\{\theta > a\}}$ for $a \in (0,1)$. For $k \in \mathbb{N}$ and $a \in (0,1)$, define $M_a(k)$ be a minimal DFA which agrees on all $x \in \{0,1\}^{<k}$ with the function $\text{MAJ}_a : \{0,1\}^* \to \{0,1\}$, defined by

$$\text{MAJ}_a(x) = \mathbb{1}_{\left\{\sum_{i=1}^{|x|} x_i > a|x|\right\}}$$

($M_{1/3}(7)$ is illustrated in Figure 4). We can associate to each $M_a(k)$ an ergodic Markov chain with a unique stationary distribution, as done in the proof of Theorem 2. Thus, each $M_a(k)$ has a well-defined limiting acceptance probability

$$\rho_a(k,\theta) = \lim_{n\to\infty} \mathbf{P}\{X^n \in M_a(k)\}$$

as well as a limiting probability of error

$$R_a(k,\theta) = \lim_{n\to\infty} \mathbf{P}\{M_a(k)(X^n) \neq T_a(\theta)\}$$

(the curves of $\rho_{1/3}(7,\cdot)$ and $\rho_{1/3}(24,\cdot)$ are plotted in Figure 5). It is not difficult to show, using arguments analogous to those in Theorem 4, that

$$\lim_{k\to\infty} R_a(k,\theta) = \begin{cases} 0, & \theta \neq a \\ \frac{1}{2}, & \theta = a \end{cases}.$$

This is a natural generalization of (19) and (20) for $a \neq \frac{1}{2}$; we leave the analysis of the convergence rates for future work.

Contrast these results with a theorem of Eisman and Ravikumar (2005), which may be stated as follows.

Theorem 7 (Eisman and Ravikumar (2005)) *Let X be a Bernoulli process with parameter $\theta = \frac{1}{2}$ and suppose that $L \subseteq \{0,1\}^*$ is a regular language. Then*

$$\limsup_{n\to\infty} \mathbf{P}\left\{L(X^n) \neq \text{MAJ}_{1/2}(X^n)\right\} \ge \frac{1}{2}.$$

Theorem 5 essentially shows that for a given Bernoulli process with parameter θ, the majority function can be ε-approximated (in the sense of (15)) by a DFA with $O\left(\frac{\log\varepsilon}{\log|1/2-\theta|}\right)$ states. Thus, (19) and (20) provide a converse of sorts to Theorem 7, which eliminates the possibility of a better than $\frac{1}{2}$ approximation to MAJ by *any* DFA under the unbiased Bernoulli process. A natural direction for generalizing Theorems 5 and 7 is the following:

Conjecture 1 Let X be a Bernoulli process with parameter θ. Then

(a) For all $\varepsilon > 0$ and $a \neq \theta$, there exists an automaton A with $O\left(\frac{\log\varepsilon}{\log|a-\theta|}\right)$ states such that

$$\limsup_{n\to\infty} \mathbf{P}\{A(X^n) \neq \text{MAJ}_a(X^n)\} \le \varepsilon.$$

(b) For $a = \theta$ and every DFA A,

$$\limsup_{n\to\infty} \mathbf{P}\{A(X^n) \neq \text{MAJ}_a(X^n)\} \ge \frac{1}{2}.$$

See Cordy and Salomaa (2007) for other results on approximating non-regular languages by DFAs.

9 Discussion and future directions

9.1 Summary

We have shown that consistent statistical estimation is not realizable by finite-state automata, but if the consistency requirement is relaxed, efficient ε-approximations exist. The negative result holds for the broad class of stationary processes with full support.

Along the way, we encountered several insights. In Section 7, we saw that although a DFA cannot accumulate statistical information, it can exploit a time drift or forbidden patterns in the random process. It would be interesting to make this intuition more rigorous — for example, by giving a full characterization of the random processes that do not admit consistent finite-state estimators of nontrivial statistical predicates.

The observations in Section 8 raise a number of interesting questions. It is natural to enquire about the optimality of the DFAs used to approximate the majority function in Theorem 5. We conjecture that any finite-state ε-approximation (see (15)) to MAJ must use $\Omega(\log(1/\varepsilon))$ states; more on this below.

9.2 Future work

Section 8 suggests a rather general technique for estimating statistical predicates by DFAs. For simplicity, let us consider the Bernoulli process $X_1, X_2, \ldots$

with parameter $\theta \in [0,1]$. For a statistical predicate $T : \theta \mapsto \{0,1\}$, we define its associated *MLE consistency set* $\Theta_T \subseteq [0,1]$ as the set of $\theta \in [0,1]$ for which

$$T(\hat{\theta}_n(X_1, \ldots, X_n)) \to T(\theta)$$

in probability, where $\hat{\theta}_n : \{0,1\}^n \to [0,1]$ is the maximum-likelihood estimator for θ:

$$\hat{\theta}_n(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

We propose to construct a finite-state estimator for T as follows. Let $M_T(k)$ a minimal DFA that agrees with $T(\hat{\theta}_n(\cdot))$ on $\{0,1\}^n$, for $n = 1, 2, \ldots, k-1$ [thanks to Daniel Berend for pointing out that such a minimal DFA need not be unique, and for debugging earlier conjectures]. For concreteness, such a DFA may be constructed via Angluin's algorithm (Angluin, 1987). Three intriguing questions immediately suggest themselves:

1. Computability: Can $M_T(k)$ be computed efficiently? A naive implementation of Angluin's algorithm requires evaluating the automaton on $\Omega(2^k)$ strings; is there a faster method? Since our goal is ε-consistent prediction, perhaps sampling could be used to efficiently construct a DFA which is ε-consistent with the "training set" $\{0,1\}^{<k}$?
2. Consistency: We conjecture that for a fixed T and all $\theta \in \Theta_T$, the sequence of automata $\{M_T(k)\}$ is $\varepsilon(k)$-consistent with $\varepsilon(k) \to 0$ as $k \to \infty$.
3. Optimality: We conjecture that the construction of $M_T(k)$ is asymptotically optimal (in terms of automaton size) among all the finite-state estimators for T, for almost all $\theta \in \Theta_T$.

Acknowledgements Ronen Brafman posed the question that motivated this whole paper. Many thanks to Gerald Eisman for debugging the numerous faulty versions of Theorem 2 and to Daniel Dadush for help with stationary distributions. Thanks to Cosma Shalizi for pointing out the papers of Cover and Hellman and to Thomas Cover, Martin Hellman, and Ron Rivest for the very gracious correspondence. I thank Daniel Berend for numerous fruitful discussions and careful readings of the manuscript, and the two anonymous reviewers for their helpful comments. Part of this work was done at the Weizmann Institute under the kind hosting and guidance of Gideon Schechtman.

References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137 – 147, 1999.

Dana Angluin. Learning regular sets from queries and counterexamples. *Inf. Comput.*, 75(2):87–106, 1987.

Ronen I. Brafman and Moshe Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* Springer-Verlag, 1999.

Amit Chakrabarti, Graham Cormode, and Andrew McGregor. Robust lower bounds for communication and stream computation. In *STOC '08: Proceedings of the 40th annual ACM symposium on Theory of computing*, 2008.

Brendan Cordy and Kai Salomaa. On the existence of regular approximations. *Theor. Comput. Sci.*, 387 (2):125–135, 2007.

Thomas M. Cover. Hypothesis testing with finite statistics. *Ann. Math. Statist.*, 40:828–835, 1969.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* Wiley-Interscience Hoboken, NJ, second edition, 2006.

Thomas M. Cover, Michael A. Freedman, and Martin E. Hellman. Optimal finite memory learning algorithms for the finite sample problem. *Information and Control*, 30(1):49–85, 1976.

Gerald Eisman and Bala Ravikumar. Approximate recognition of non-regular languages by finite automata. In *ACSC '05: Proceedings of the Twenty-eighth Australasian conference on Computer Science*, pages 219–227, 2005.

Joan Feigenbaum, Sampath Kannan, Martin Strauss, and Mahesh Viswanathan. An approximate l1-difference algorithm for massive data streams. *SIAM J. Comput.*, 32(1):131–151, 2002.

Philippe Flajolet and G. Nigel Martin. Probabilistic counting algorithms for data base applications. *J. Comput. System Sci.*, 31(2):182–209, 1985.

Sudipto Guha and Andrew McGregor. Stream order and order statistics: Quantile estimation in random-order streams. *SIAM J. Comput.*, 38(5):2044–2059, 2009.

Martin E. Hellman and Thomas M. Cover. Learning with finite memory. *Ann. Math. Statist*, 41:765–782, 1970.

Martin E. Hellman and Thomas M. Cover. On memory saved by randomization. *Ann. Math. Statist.*, 42: 1075–1078, 1971.

M.E. Hellman. The effects of randomization on finite memory decision schemes. In *Adaptive Processes (9th) Decision and Control, 1970. 1970 IEEE Symposium on*, volume 9, pages 32 –32, 7-9 1970.

Monika R. Henzinger, Prabhakar Raghavan, and Sridhar Rajagopalan. Computing on data streams. pages 107–118, 1999.

Patrick Hirschler and Thomas M. Cover. A finite memory test of the irrationality of the parameter of a coin. *Ann. Statist.*, 3(4):939–946, 1975.

Patrick Robert Hirschler. Finite memory algorithms for testing Bernoulli random variables. *Information and Control*, 24:11–19, 1974.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58:13–30, 1963.

Piotr Indyk and David Woodruff. Optimal approximations of the frequency moments of data streams. In *STOC '05: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 202–208, New York, NY, USA, 2005. ACM.

Olav Kallenberg. *Foundations of modern probability. Second edition. Probability and its Applications.* Springer-Verlag, 2002.

John G. Kemeny and J. Laurie Snell. *Finite Markov chains.* Springer-Verlag, New York, 1976.

Leonid Kontorovich and Kavita Ramanan. Concentration Inequalities for Dependent Random Variables via the Martingale Method. *Annals of Probability*, 36(6):2126–2158, 2008.

K. B. Lakshmanan and B. Chandrasekaran. Compound hypothesis testing with finite memory. *Information and Control*, 40(2):223–233, 1979.

Frank Thomson Leighton and Ronald L. Rivest. Estimating a probability using finite memory. In *Estimating a Probability using Finite Memory*, *IEEE Trans. Inform. Theory*, **IT-32**, 733–742, 1986.

Harry R. Lewis and Christos H. Papadimitriou. *Elements of the Theory of Computation.* Prentice Hall, 1981.

Robert Morris. Counting large numbers of events in small registers. *Commun. ACM*, 21(10):840–842, 1978.

S. Muthukrishnan. Data streams: algorithms and applications. *Found. Trends Theor. Comput. Sci.*, 1(2): 117–236, 2005.

Doron Rajwan and Meir Feder. Universal finite memory machines for coding binary sequences. In *Data Compression Conference*, pages 113–122, 2000.

F. Samaniego. Estimating a binomial parameter with finite memory. *Information Theory, IEEE Transactions on*, 19(5):636 – 643, sep 1973.

Michael Sipser. *Introduction to the Theory of Computation.* Course Technology, 2005.

Boris A. Trakhtenbrot and Janis M. Barzdin'. *Finite Automata: Behavior and Synthesis*, volume 1 of *Fundamental Studies in Computer Science.* North-Holland, Amsterdam, 1973.