RC22953 (W0310-200) October 31, 2003
Mathematics

IBM Research Report

Non-Asymptotic Upper Bounds on the Probability of the 0-Atypical Set for Markov Models

Luis Alfonso Lastras-Montaño
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Non-Asymptotic Upper Bounds on the Probability of the ϵ-atypical set for Markov chains

Luis Alfonso Lastras-Montaño *

October 31, 2003

Abstract

For an irreducible, aperiodic finite alphabet Markov chain, consider the set of sequences with atypical probability $A_{l,\epsilon} \triangleq \{\mathbf{x}_0^{l-1} : \left|-l^{-1}\log_2 P(\mathbf{x}_0^{l-1}) - h\right| > \epsilon\}$. Well known results demonstrate that for every $\epsilon > 0$, $\lim_{l\to\infty} -l^{-1}\log_2 P(A_{l,\epsilon})$ is positive, thus the probability of the ϵ-atypical set decays exponentially fast for sufficiently large values of l. There are many practical situations in which we require good bounds that hold when the parameter ϵ is allowed to vary with l in a manner relevant to the problem in question. In this correspondence we derive bounds with this property based on the method of Markov types of Davisson, Longo and Sgarro.

1 Introduction

At the heart of many information theory results lies the notion of entropy typicality, the usefulness of which is validated by Shannon's Asymptotic Equipartition Property. Said property is often cited in its weakened probability form: for stationary and ergodic sources and for fixed $\epsilon > 0$, the probability of the set

$$A_{\epsilon,l} = \left\{\mathbf{x}_0^{l-1} : \left|-l^{-1}\log_2 P(\mathbf{x}_0^{l-1}) - h\right| > \epsilon\right\} \tag{1}$$

can be made as close to zero as desired by making l large enough.

Consider a finite alphabet Markov chain $X_0, X_1, \ldots$ that is both irreducible and aperiodic with transition probability matrix P, stationary distribution π and finite cardinality alphabet $\mathcal{Q} = \{1, 2, \cdots, |\mathcal{Q}|\}$. In some instances, it is of interest to gain insight on the rate at which $P(A_{\epsilon,l})$ decays to zero as $l \to \infty$ for a given ϵ_l-sequence satisfying $\epsilon_l \to 0$. If ϵ_l decays too fast, then $P(A_{l,\epsilon_l})$ may not tend to zero as $l \to \infty$; on the other hand in the case where $\epsilon_l = \epsilon$ well known arguments can be used to demonstrate that $P(A_{l,\epsilon_l})$ decays exponentially fast with l.

It is of practical interest to investigate situations intermediate to the ones mentioned above; for example see the recent work on the redundancy of the sliding-window Lempel-Ziv algorithm [1]. The method of types [2] can be used to treat the iid case effectively by starting from the inequality [3]

$$P\left(\left\{\mathbf{x}_0^{l-1} : \sum_{j=1}^{|\mathcal{Q}|}\left|\hat{p}(j|\mathbf{x}_0^{l-1}) - \mu_1(a_j)\right| \geq \epsilon\right\}\right) \leq (l+1)^{|\mathcal{Q}|}2^{-lc\epsilon^2} \tag{2}$$

where $\hat{p}(\cdot|\mathbf{x}_0^{l-1}) = N[\cdot|\mathbf{x}_0^{l-1}]/l$ denotes the empirical probability mass function (pmf) associated with $\mathbf{x}_0^{l-1}$, $\mu_1(\cdot)$ is the true pmf and c is a positive constant. Inequality (2) is true for every $l > 0$ and $\epsilon > 0$ and therefore can be used to treat the problem considered here.

*Luis Lastras is with the Server Technology Department, IBM TJ Watson Research Center, Yorktown Heights, NY 10598

In the Markov case the situation does not appear to be as straightforward since a sampling of the literature [4] [3] revealed that all relevant results assume a fixed value for ϵ and then require that l be large enough, where the notion of how large is "large" is essentially hidden from the reader for most practical purposes.

In this correspondence we address this gap by building upon the notion of Markov type of Davisson, Longo and Sgarro [5]. While the exact notion of a Markov type will be discussed in the next section, we immediately present our main result:

Theorem 1 *There exist constants $\kappa_1(P) > 0$ and $\kappa_2(P) > 0$ such that for every pair (ϵ, l) that satisfies $l - 1 > \kappa_2(P)/\epsilon$ and for every ϵ-atypical type $\mathcal{T}$,*

$$P(\mathcal{T}) \leq 2^{-(l-1)\kappa_1(P)\epsilon^2}$$

We stress that one of the positive features of the proof of this result is that the constants involved can be computed direclty from P in a simple manner, this is, we have not put forth an "only existence" result. Moreover the condition $l - 1 > \kappa_2(P)/\epsilon$ is mild by most standards, stated in words the only requirement is that the form of the decay of ϵ_l cannot be as fast as $1/l$. Finally note that every Markov model where the memory depth is greater than one can be thought of as a first order Markov chain with appropriate enlargement of the alphabet, and thus these results also apply to that case.

The probability of the ϵ-atypical set is no larger than $|\mathcal{Q}| \left(1 + \frac{l-1}{|\mathcal{Q}|}\right)^{|\mathcal{Q}|^2}$ times the right hand side in the Theorem; note that unlike frequent uses of the method of types, good estimates for the number of different types are essential in problems of this nature (for this estimate see the arguments in the Preliminaries).

Better but more complex bounds may be obtained by means of the following method: we may label types as "atypical" or "very atypical". Very atypical types are ξ-atypical, where $\xi > 0$ is a constant that is independent of the sequence length. One then may use standard arguments to show that the probability of the union of the very atypical types will decay exponentially fast with an exponent that is essentially as good as possible. Atypical types are ϵ_l-atypical but not very atypical; the trick is then to count the number of atypical types and to choose the threshold ξ judiciously. We do not explore these improved bounds in this correspondence.

1.1 Preliminaries

As previewed, our first task is to provide a suitably embodiment to the idea of type that is relevant to the present situation. Every statement in these Preliminaries is part of the standard method of types machinery; we take advantage of this opportunity to introduce notation and to point out a few subtleties that are relevant to the main resulit. We denote by $N[(j,k)|\mathbf{x}_0^{l-1}]$ the number of times the pair (j,k) appears in the l-sequence $\mathbf{x}_0^{l-1}$ and by $N[j|\mathbf{x}_0^{l-1}]$ the number times that j appears in $\mathbf{x}_0^{l-1}$. The tuple

$$\left(\mathbf{x}_0, N[\cdot|\mathbf{x}_0^{l-1}]\right) \tag{3}$$

is called the Markov type (and subsequently, simply the "type") of the sequence $\mathbf{x}_0^{l-1}$; the reason the starting symbol is included will be revealed in the subsequent discussion. The set of sequences that share the same type is called a type class. We will use the literal $\mathcal{T}$ to refer to a type *and* to the set of all sequences that share it. Note that

$$N[j|\mathbf{x}_0^{l-2}] \;=\; \sum_k N[(j,k)|\mathbf{x}_0^{l-1}], \qquad \sum_j N[j|\mathbf{x}_0^{l-2}] = l - 1$$

and that as a consequence, the $|\mathcal{Q}|$ integers in $N[(j,\cdot)|\mathbf{x}_0^{l-1}]$ can be chosen in at most $\left(1+N[j|\mathbf{x}_1^{l-1}]\right)^{|\mathcal{Q}|}$ ways. Forming the product of the resulting $|\mathcal{Q}|$ bounds, applying exp log, using Jensen's inequality and noting that there are $|\mathcal{Q}|$ ways of selecting the first symbol yields the estimate used in the Introduction.

Let $\mathcal{T}$ be a given type for l-sequences and let $\mathbf{z}_0^{l-1}$ be any sequence in $\mathcal{T}$. If the tuple (j,k) appears at least once in $\mathbf{z}_0^{l-1}$ and $P_{k|j}=0$ then it is trivial to see that $P(\mathcal{T})=0$. Therefore, we shall assume that

$$\text{if } \; P_{k|j}=0 \text{ then } \; N[(j,k)|\mathbf{z}_0^{l-1}]=0 \tag{4}$$

Define the $|\mathcal{Q}|\times|\mathcal{Q}|$ matrix $\hat{P}$ and the $|\mathcal{Q}|\times 1$ vector $\hat{\pi}$ as

$$\hat{P}_{k|j} \;\triangleq\; \begin{cases} \frac{N[(j,k)|\mathbf{z}_0^{l-1}]}{N[j|\mathbf{z}_0^{l-2}]} & \text{if } N[j|\mathbf{z}_0^{l-2}]>0 \\ P_{k|j} & \text{otherwise} \end{cases} \tag{5}$$

$$\hat{\pi}_j \;\triangleq\; \frac{N[j|\mathbf{z}_0^{l-2}]}{l-1} \tag{6}$$

It is clear that Definitions (5) and (6) are independent of which sequence $\mathbf{z}_0^{l-1}\in\mathcal{T}$ was selected. Also, it is not difficult to verify that $\hat{P}$ is a stochastic matrix and that the entries of $\hat{\pi}$ define a probability mass function. Combining Assumption (4) and Definition (5) it can be easily seen that

$$\text{if } \; P_{k|j}=0 \;\; \text{ then } \;\; \hat{P}_{k|j}=0, \tag{7}$$

a remark that will find its use shortly. Although in general it is not true that $\hat{P}\hat{\pi}=\hat{\pi}$, one can derive a similar useful statement. Note that the type $\mathcal{T}$ must fall in one of these two categories:

(a) Every sequence in $\mathcal{T}$ starts and ends with the same symbol, and

(b) There exist two symbols $j_{\mathcal{T}}$ and $k_{\mathcal{T}}$ such that every sequence in $\mathcal{T}$ starts with $j_{\mathcal{T}}$ and ends with $k_{\mathcal{T}}$

If the type satisfies (a) then $\hat{P}\hat{\pi}=\hat{\pi}$; here we define $e_{\mathcal{T}}$ to be a null $|\mathcal{Q}|\times 1$ vector. In case (b), we have that $\hat{P}\hat{\pi}\neq\hat{\pi}$ and we define $e_{\mathcal{T}}$ to be a vector of zeros except at position $k_{\mathcal{T}}$, where it is equal to $1/(l-1)$, and position $j_{\mathcal{T}}$, where it is equal to $-1/(l-1)$. Then the following is true for every type $\mathcal{T}$:

$$\begin{aligned} \hat{P}\hat{\pi} &= \hat{\pi}+e_{\mathcal{T}} \\ \|e_{\mathcal{T}}\|_1 &\leq 2/(l-1) \end{aligned} \tag{8}$$

Of fundamental interest is to have good bounds for the cardinality of a type class and its probability as governed by a given Markov chain. Let Q any $|\mathcal{Q}|\times|\mathcal{Q}|$ stochastic matrix that safisfies

$$\text{if } Q_{k|j}=0 \;\; \text{ then } \;\; \hat{P}_{k|j}=0$$

examples of which include $\hat{P}$ itself and P, as evidenced by (7). Assumption (4) implies that for every sequence $\mathbf{x}_0^{l-1}\in\mathcal{T}$, $\Pi_{i=1}^{l-1}Q_{\mathbf{x}_i|\mathbf{x}_{i-1}}>0$ and therefore it is feasible to write

$$\begin{aligned} -\frac{1}{l-1}\log_2\left(\Pi_{i=1}^{l-1}Q_{\mathbf{x}_i|\mathbf{x}_{i-1}}\right) &= -\sum_{(j,k):Q_{k|j}>0}\frac{N[(j,k)|\mathbf{z}_0^{l-1}]}{l-1}\log_2 Q_{k|j} \\ &= -\sum_{(j,k):\hat{P}_{k|j}>0}\hat{\pi}_j\hat{P}_{k|j}\log_2 Q_{k|j} \\ &\triangleq H(\hat{P})+D(\hat{P}\|Q) \end{aligned} \tag{9}$$

where

$$\begin{aligned} H(\hat{P}) &= -\sum_{j=1}^{|\mathcal{Q}|} \hat{\pi}_j \sum_{k:\hat{P}_{k|j}>0} \hat{P}_{k|j} \log_2 \hat{P}_{k|j} \\ D(\hat{P}\|Q) &= \sum_{j=1}^{|\mathcal{Q}|} \hat{\pi}_j \sum_{k:\hat{P}_{k|j}>0} \hat{P}_{k|j} \log_2 \frac{\hat{P}_{k|j}}{Q_{k|j}} \end{aligned}$$

Setting $Q = \hat{P}$ yields

$$\hat{P}\left(\mathbf{X}_1^{l-1} = \mathbf{x}_1^{l-1} | \mathbf{X}_0 = \mathbf{z}_0\right) = 2^{-(l-1)H(\hat{P})} \qquad \forall\ \mathbf{x}_0^{l-1} \in \mathcal{T}$$

From this it is straightforward to see that $|\mathcal{T}|2^{-(l-1)\hat{H}} = \hat{P}\left(\mathcal{T}|\mathbf{X}_0 = \mathbf{z}_0\right) \leq 1$ thus yielding the bound $|\mathcal{T}| \leq 2^{(l-1)\hat{H}}$. Now setting $Q = P$ in (9) and using this bound we obtain

$$P(\mathcal{T}) \leq P\left(\mathcal{T}|\mathbf{X}_0 = \mathbf{z}_0\right) \leq 2^{-(l-1)D(\hat{P}\|P)} \tag{10}$$

Thus one can upper bound $P\left(\mathcal{T}\right)$ by lower bounding $D(\hat{P}\|P)$.

1.2 Results

Theorem 1 (restated) *There exist constants $\kappa_1(P) > 0$ and $\kappa_2(P) > 0$ such that for every pair (ϵ, l) that satisfies $l - 1 > \kappa_2(P)/\epsilon$ and for every ϵ-atypical type $\mathcal{T}$,*

$$P(\mathcal{T}) \leq 2^{-(l-1)\kappa_1(P)\epsilon^2}$$

Proof. Let $\hat{P}$ and $\hat{\pi}$ be the empirical conditional probabilities and marginal associated with type $\mathcal{T}$. Assume that $\mathbf{x}_0$ is the symbol that every sequence in $\mathcal{T}$ starts with. Set $\Delta = P - \hat{P}$. Let $|\Delta|$ be the $|\mathcal{Q}| \times |\mathcal{Q}|$ matrix with (k, j) entry equal to the absolute value of the corresponding entry in the Δ matrix, this is, $|\Delta|_{k,j} = |\Delta_{k,j}|$. Since $\mathcal{T}$ is ϵ-atypical,

$$\begin{aligned} \epsilon &< \left| -\sum_j \sum_{k:P_{k|j}>0} (\pi_j P_{k|j} - \hat{\pi}_j \hat{P}_{k|j}) \log_2 P_{k|j} \right| \\ &= \left| -\sum_j \sum_{k:P_{k|j}>0} (\pi_j P_{k|j} - \hat{\pi}_j P_{k|j} + \hat{\pi}_j P_{k|j} - \hat{\pi}_j \hat{P}_{k|j}) \log_2 P_{k|j} \right| \\ &= \left| -\sum_j (\pi_j - \hat{\pi}_j) \sum_{k:P_{k|j}>0} P_{k|j} \log_2 P_{k|j} + \sum_j \hat{\pi}_j \sum_{k:P_{k|j}>0} \Delta_{k,j} \log_2 P_{k|j} \right| \\ &\leq \|\pi - \hat{\pi}\|_1 \max_j H(P(\cdot|j)) + \||\Delta|\hat{\pi}\|_1 \left(\max_{j,k:P_{k|j}>0} -\log_2 P_{k|j} \right) \\ &\leq \left(\kappa_3^{-1}(P) \max_j H(P(\cdot|j)) + \left(\max_{j,k:P_{k|j}>0} -\log_2 P_{k|j} \right) \right) \||\Delta|\hat{\pi}\|_1 \\ &\quad + 2\kappa_3^{-1}(P) \max_j H(P(\cdot|j))/(l-1) \end{aligned} \tag{11}$$

where the last inequality follows from Lemma 1 which is proved below. Somewhat arbitrarily and in order to simplify the shape of the bound, assume that

$$\begin{aligned} l-1 &> \kappa_2(P)/\epsilon \\ \kappa_2(P) &= 4\kappa_3^{-1}(P)\max_j H(P(\cdot|j)) \end{aligned}$$

Then (11) translates into

$$\epsilon\sqrt{\kappa_1(P)2\ln 2} < \||\Delta|\hat{\pi}\|_1 \tag{12}$$

defining $\kappa_1(P)$ appropriately; note that we end up with a postive constant as demanded. By Pinsker's inequality,

$$\begin{aligned} D(\hat{P}\|P) &= \sum_{j=1}^{|\mathcal{Q}|}\hat{\pi}_j \sum_{k:\hat{P}_{k|j}>0}\hat{P}_{k|j}\log_2\frac{\hat{P}_{k|j}}{P_{k|j}} \\ &= \sum_{j=1}^{|\mathcal{Q}|}\hat{\pi}_j \sum_{k:P_{k|j}>0}\hat{P}_{k|j}\log_2\frac{\hat{P}_{k|j}}{P_{k|j}} \\ &\geq \frac{1}{2\ln 2}\sum_{j=1}^{|\mathcal{Q}|}\hat{\pi}_j\left(\sum_{k:P_{k|j}>0}|\Delta_{k,j}|\right)^2 \\ &\overset{(a)}{\geq} \frac{1}{2\ln 2}\left(\sum_{j=1}^{|\mathcal{Q}|}\sum_{k:P_{k|j}>0}^{|\mathcal{Q}|}\hat{\pi}_j|\Delta_{k,j}|\right)^2 \\ &= \frac{1}{2\ln 2}\||\Delta|\hat{\pi}\|_1^2 \end{aligned} \tag{13}$$

where (a) follows from Jensen's inequality. Combining (12) with (13) ends the proof. ———————

We now state and proof the Lemma used in Theorem 1. Note in addition to the application above, this Lemma can also be used to obtain "non-asymptotic" large deviations bounds for the empirical marginal.

Lemma 1 *There exists a constant $\kappa_3(P) > 0$ such that $\kappa_3(P)\|\pi-\hat{\pi}\|_1 \leq \||\Delta|\hat{\pi}\|_1 + 2/(l-1)$*

Proof. Let $\delta = \pi - \hat{\pi}$. In light of $\||\Delta|\hat{\pi}\|_1 > \|\Delta\hat{\pi}\|_1$, our goal will be to obtain a lower bound for $\|\Delta\hat{\pi}\|_1$ in terms of $\|\delta\|_1$. Substituting the definitions of Δ and δ in the equality $P\pi = \pi$, we obtain

$$(\hat{P}+\Delta)(\hat{\pi}+\delta) = \hat{\pi}+\delta \tag{14}$$

Using (8) we obtain

$$e_{\mathcal{T}} + \Delta\hat{\pi} = (I-P)\delta$$

Application of the L_1 norm and the triangle inequality yields

$$2(l-1)^{-1} + \|\Delta\hat{\pi}\|_1 \geq \|(I-P)\delta\|_1$$

Since $\|P\|_1 = 1$ we have that for every $k \geq 1$,

$$\begin{aligned} \|(I - P^k)\delta\|_1 &\geq \|P(I - P^k)\delta\|_1 \\ &= \|(I - P^{k+1})\delta - (I - P)\delta\|_1 \\ &\geq \|(I - P^{k+1})\delta\|_1 - \|(I - P)\delta\|_1 \end{aligned}$$

and therefore for every positive integer n,

$$\begin{aligned} \|(I - P)\delta\|_1 &\geq \frac{\|(I - P^n)\delta\|_1}{n} \\ &\geq \frac{\|\delta\|_1 - \|P^n\delta\|_1}{n} \end{aligned} \tag{15}$$

Now

$$\begin{aligned} \|P^n\delta\|_1 &= \sum_{k=1}^{|\mathcal{Q}|} \left| \sum_{j=1}^{|\mathcal{Q}|} \left(P^n_{k|j} - \pi_k + \pi_k \right) \delta_j \right| \\ &\overset{(a)}{=} \sum_{k=1}^{|\mathcal{Q}|} \left| \sum_{j=1}^{|\mathcal{Q}|} \left(P^n_{k|j} - \pi_k \right) \delta_j \right| \\ &\leq \sum_{k=1}^{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \left| P^n_{k|j} - \pi_k \right| |\delta_j| \\ &\leq \|\delta\|_1 |\mathcal{Q}| \max_{j,k} \left| P^n_{k|j} - \pi_k \right| \end{aligned}$$

where (a) follows from the fact that the sum of the entries of δ is equal to zero. Assume that all the entries of the matrix P are positive. Since the state space is finite and the chain is assumed to be irreducible and aperiodic [6],

$$\left| P^n_{k|j} - \pi_k \right| \leq \left(1 - |\mathcal{Q}| \min_{j,k} P_{k|j} \right)^n \tag{16}$$

thereby giving

$$\inf_{\delta : 1^T\delta = 0, \delta \neq 0} \frac{\|(I - P)\delta\|_1}{\|\delta\|_1} \geq \frac{1 - |\mathcal{Q}| \left(1 - |\mathcal{Q}| \min_{j,k} P_{k|j}\right)^n}{n} \triangleq \kappa_3(P)$$

It is clear that there is a first n such that the right hand side will be striclty positive; we give no further indication as to how to choose the value of n that will give the largest lower bound. If the matrix $P_{k|j}$ has some entries equal to zero, then we proceed as follows: since P is irreducible and aperiodic, then P^m has all positive entries for some m. The constant now becomes

$$\kappa_3(P) = \frac{1 - |\mathcal{Q}| \left(1 - |\mathcal{Q}| \min_{j,k} P^m_{k|j}\right)^{mn}}{mn}$$

References

[1] Luis A. Lastras-Montaño. On the Redundancy of the Sliding Window Lempel-Ziv Algorithm. *In preparation.*

[2] I. Csiszár, J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.

[3] P. C. Shields. *The Ergodic Theory of Discrete Sample Paths.* American Mathematical Society, 1996.

[4] Ofer Zeitouni Amir Dembo. *Large Deviations Techniques And Applications.* Springer, 1998.

[5] Lee D. Davisson, Giuseppe Longo, Andrea Sgarro. The Error Exponent for the Noiseless Encoding of Finite Ergodic Markov Sources. *IEEE Trans. Inform. Theory*, 27(4):431–438, July 1994.

[6] Patrick Billingsley. *Probability and Measure.* John Wiley and Sons, 1995.