ISIT'2013 1569720261

Estimation in slow mixing, long memory channels

Meysam Asadi* Email: masadi@hawaii.edu
Ramezan Paravi Torghabeh* Email: paravi@hawaii.edu
Narayana P. Santhanam* Email: nsanthan@hawaii.edu

***Abstract*—We consider estimation of binary channels with memory where the transition probabilities (*channel parameters*) from the input to output are determined by prior outputs (*state of the channel*). While the channel is unknown, we observe the joint input/output process of the channel—we have n *i.i.d.* input bits and their corresponding outputs. Motivated by applications related to the backplane channel, we want to estimate the channel parameters as well as the stationary probabilities for each state.**

Two distinct problems complicate estimation in this setting: (i) long memory, and (ii) slow mixing which could happen even with only one bit of memory. In this setting, any consistent estimator can only converge pointwise over the model class. Namely, given any estimator and any sample size n, the underlying model could be such that the estimator performs poorly on a sample of size n with high probability. But can we look at a length-n sample and identify *if* an estimate is likely to be accurate?

Since the memory is unknown a-priori, a natural approach, known to be consistent, is to estimate a potentially coarser model with memory $k_n = \alpha_n \log n$, where α_n is a function that grows $\mathcal{O}(1)$. Note however that (i) the coarser model is estimated using only samples from the true model; and (ii) we want the best possible answers with a length-n sample, rather than just consistency. Combining results on universal compression and Aldous' coupling arguments, we obtain sufficient conditions (even for slow mixing models) to identify when naive (i) estimates of the channel parameters and (ii) estimates related to the stationary probabilities of the channel states are accurate, and bound their deviations from true values.

I. Introduction

Universal compression and estimation often go hand in hand, at least in the *i.i.d.* setting. In the (*i.i.d.*) universal compression setting, we have a class of distributions $\mathcal{P}$ over a discrete set $\mathcal{A}$, and we obtain a sample in $\mathcal{A}^n$ (the set of length-n strings of symbols from $\mathcal{A}$) via n *i.i.d.* draws from some (unknown) distribution $p \in \mathcal{P}$. Assuming $\mathcal{P}$ has worst-case redundancy that grows sublinearly in n, we can often leverage a universal distribution q for $\mathcal{P}$ to *estimate* the unknown distribution in $\mathcal{P}$—for example, the universal KT or Laplace estimator for the class of *i.i.d.* binary distributions.

However, when we introduce memory into the picture, the above picture gets muddied. Consider a length-n sample obtained from an unknown source in the class of binary Markov sources with memory one. No matter what the source is, the context tree weighting algorithm can give the sample a codelength that is at most the true codelength plus $\mathcal{O}(\log n)$—namely it does not underestimate the true probability by more than a subexponential factor as n increases. However, as we will see, irrespective of how large the sample size n is, we may be unable to estimate the stationary probabilities of 0s and 1s reliably.

Let the transition probability from 1 to 0 in our memory-1 source be $\epsilon \ll 1/n$. By picking the transition probability from 0 to 1 appropriately, we can set the stationary probabilities of 1s and 0s in a wide range, without changing how a length-n sample will look like.

Example 2 gives two such sources with stationary probabilities (1/2,1/2) and (2/3,1/3). Now, if we start from 1, either source will, with high probability, yield a sequence of n 1s, or perhaps a long string of 1s with a few 0s bunched together at the end depending on the value of ϵ. Say, for the sake of a concrete example, that in this sample $\mathbf{x}_1$, we have $n - \log n$ 1s followed by a string of $\log n$ 0s. Thus it is not possible to estimate stationary probabilities from the sample $\mathbf{x}_1$. This particular phenomenon, where the number of times each state (0 and 1 here) appears is very different from their stationary probabilities is often formalized as *slow mixing*, see [1].

Estimation for Markov processes has been extensively studied, see for example [2]–[8]. In these papers, the authors have considered (i) consistency of estimators, (ii) bounds on estimates that hold eventually almost surely, and (iii) bounds that hold for all sample sizes but which depend on the model parameters. Limit theorems that hold eventually almost surely for relative frequencies of finite length blocks of Markov chains with arbitrary order are proved in [9], [10].

In this paper, we deal with stationary ergodic Markov processes with a finite alphabet and are motivated by a channel estimation problem we will describe shortly. We emphasize at the outset that we do *not* exclude slow mixing processes. In fact, our philosophy will be: given n samples, what is the best we can say, if anything?

In other words, we have an estimation problem where any estimator can only converge pointwise to the true values, rather than uniformly, over the model class. Yet, we can still ask a useful question—can we look at the data and say if we are doing well? Contrast the above sample $\mathbf{x}_1$ with a new sample $\mathbf{x}_2$, also with $n - \log n$ 1s and $\log n$ 0s, but $\mathbf{x}_2$ has 0s spread uniformly in the sequence. Unlike in the case of $\mathbf{x}_1$, upon seeing $\mathbf{x}_2$ we may want to conclude that we have an *i.i.d.* source with a high probability for 1.

The particular application we are motivated by arises in high speed chip-to-chip communications, and is commonly called the backplane channel [11]. Here, residual reflections between inter-chip connects form a significant source of interference. Because of parasitic capacitances, the channel is

This work was supported by NSF Grants CCF-1065632, CCF-1018984 and EECS-1029081. We thank A. Kavčić for helpful discussions. * The authors are with the Department of Electrical Engineering, University of Hawai'i at Mānoa, Honolulu, HI.

highly non-linear as well, and consequently the residual signal that determines the channel state is not a linear function of past inputs as in typical interference channels. We therefore consider a channel model where the output is not necessarily a linear function of the input, and in addition, the channel encountered by the k'th input bit is determined by the prior outputs. Therefore, we begin with estimation problems in channels whose state is determined by the output memory.

Our main results are in Theorems 3-4. These results show how to look at a data sample and identify states of the channel that are amenable to accurate estimation from the sample. They also allow us to sometimes (depending on how the data looks) conclude that certain naive estimators of stationary probabilities or channel transition probabilities happen to be accurate, *even if the channel evolution is slow mixing*. To obtain these results, we combine universal compression results of the context tree weighting algorithm with coupling arguments by Aldous [12]. While most proofs are omitted in this extended abstract, they are available from [13]. Furthermore, these results throw up several new questions, see [13].

II. Markov Processes and Channels

a) Alphabet and strings: Most notation here is standard, we include them for completeness. $\mathcal{A}$ is a finite alphabet with cardinality $|\mathcal{A}|$, $\mathcal{A}^* = \bigcup_{k\geq 0}\mathcal{A}^k$ and $\mathcal{A}^\infty$ denotes the set of all semi-infinite strings of symbols in $\mathcal{A}$. We denote the length of a string $\mathbf{u} = u_1,\ldots,u_l \in \mathcal{A}^l$ by $|\mathbf{u}|$, and use $\mathbf{u}_i^j = (u_i,\cdots,u_j)$. The concatenation of strings $\mathbf{w}$ and $\mathbf{v}$ is denoted by $\mathbf{wv}$. A string $\mathbf{v}$ is a *suffix* of $\mathbf{u}$, denoted by $\mathbf{v} \preceq \mathbf{u}$, if there exists a string $\mathbf{w}$ such that $\mathbf{u} = \mathbf{wv}$. A set $\mathcal{T}$ of strings is *suffix-free* if no string of $\mathcal{T}$ is a suffix of any other string in $\mathcal{T}$.

b) Trees: As in [14] for example, we use full binary trees to represent the states of a Markov process. We denote full trees $\mathcal{T}$ as a suffix-free set $\mathcal{T} \subset \mathcal{A}^*$ of strings (the *leaves*) whose lengths satisfy Kraft's lemma with equality. The depth of the tree $\mathcal{T}$ is defined as $\kappa(\mathcal{T}) = max\{\,|\mathbf{u}| \;:\; \mathbf{u}\in\mathcal{T}\,\}$. A string $\mathbf{v}\in\mathcal{A}^*$ is an *internal node* of $\mathcal{T}$ if either $\mathbf{v}\in\mathcal{T}$ or there exists $\mathbf{u}\in\mathcal{T}$ such that $\mathbf{v}\preceq\mathbf{u}$. The *children* of an internal node $\mathbf{v}$ in $\mathcal{T}$, are those strings (if any) $a\mathbf{v}, a\in\mathcal{A}$ which are themselves either internal nodes or leaves in $\mathcal{T}$. For any internal node $\mathbf{w}$ of a tree $\mathcal{T}$, let $\mathcal{T}_\mathbf{w} = \{\mathbf{u}\in\mathcal{T} : \mathbf{w}\preceq\mathbf{u}\}$ be the subtree rooted at $\mathbf{w}$. Given two trees $\mathcal{T}_1$ and $\mathcal{T}_2$, we say that $\mathcal{T}_1$ is included in $\mathcal{T}_2$ ($\mathcal{T}_1\preceq\mathcal{T}_2$), if all the leaves in $\mathcal{T}_1$ are either leaves or internal nodes of $\mathcal{T}_2$.

c) Models: Let $\mathcal{P}^+(\mathcal{A})$ be the set of all probability distributions on $\mathcal{A}$ such that every probability is strictly positive.

Definition 1. A context tree *model* is a finite full tree $\mathcal{T}\subset\mathcal{A}^*$ with a set of probability distributions $q_\mathbf{s}\in\mathcal{P}^+(\mathcal{A})$ assigned to each $\mathbf{s}\in\mathcal{T}$. We will refer to the elements of $\mathcal{T}$ as *states* (*contexts*) and $q(\mathcal{T}) = \{q_\mathbf{s}(a) : \mathbf{s}\in\mathcal{T}, a\in\mathcal{A}\}$ as the set of *state transition probabilities* or *process parameters*. □

Every model $(\mathcal{T}, q(\mathcal{T}))$ allows for an irreducible, aperiodic[1] and ergodic [15] Markov process. Such Markov process has a unique stationary distribution μ satisfying $\mu Q = \mu$, where Q is the standard transition probability matrix formed using $q(\mathcal{T})$. Let $p_{\mathcal{T},q}$ be the unique stationary Markov process $\{\ldots,Y_0,Y_1,Y_2,\ldots\}$ which takes values in $\mathcal{A}$ satisfying

$$p_{\mathcal{T},q}(Y_1|Y_{-\infty}^0) = q_\mathbf{s}(Y_1),$$

whenever $\mathbf{s} = \mathbf{c}_\mathcal{T}(Y_{-\infty}^0)$, where $\mathbf{c}_\mathcal{T} : \mathcal{A}^\infty \to \mathcal{T}$ assigns the unique suffix $\mathbf{s}\preceq Y_{-\infty}^0$ in $\mathcal{T}$. As a note, when we write out actual strings in transition probabilities as in $q_{1000}(0)$, the state 1000 is the sequence of bits as we encounter them when reading the string left to right. If 0 follows $\cdots 1100$, the next state is a suffix of $\cdots 11000$, and if 1 follows $\cdots 1100$, the next state is a suffix of $\cdots 11001$.

Observation 1. Note that any model $(\mathcal{T}, q(\mathcal{T}))$ yields the same Markov process as a model $(\mathcal{T}', q'(\mathcal{T}'))$ where $\mathcal{T}\preceq\mathcal{T}'$ and for all $\mathbf{s}'\in\mathcal{T}'$, $q'_{\mathbf{s}'}(\cdot) = q_{\mathbf{c}_\mathcal{T}(\mathbf{s}')}(\cdot)$. □

A. Channel Model

We focus on Markov channels defined as follows. Both input $\{X_i\}_{i\geq 1}$ and output $\{Y_i\}_{i\geq 1}$ are finite alphabet processes taking values in $\mathcal{A}$ and the state of channel in each instant depend on sequence of prior outputs of the channel. The input process is drawn from an *i.i.d. Bernoulli* process, namely $P(X_i = a) = p_a$ for all $i\in\mathbb{N}$ and $a\in\mathcal{A}$, provided that $\sum_{a\in\mathcal{A}} p_a = 1$. We assume that there is no feedback in this channel setup. The joint probability distribution of the channel factorizes as

$$P(x_1^n, y_1^n|y_{-\infty}^0) = \prod_{j=1}^n P(x_j)P(y_j|y_{-\infty}^{j-1}, x_j). \tag{1}$$

The *state* of the channel at time j is therefore determined by $y_{-\infty}^{j-1}$. We consider finite memory channels, and model the possible states of the channels as leaves of a finite full binary tree, $\mathcal{T}$. Recall that $\mathbf{c}_\mathcal{T}(y_{-\infty}^{j-1})\in\mathcal{T}$ is the unique $\mathbf{s}\in\mathcal{T}$ such that $\mathbf{s}\preceq y_{-\infty}^{j-1}$. Therefore, we obtain

$$P(x_1^n, y_1^n|y_{-\infty}^0) = \prod_{j=1}^n P(x_j)P(y_j|\mathbf{c}_\mathcal{T}(y_{-\infty}^{j-1}), x_j).$$

Then $\{(X_i,Y_i)\}_{i\geq 1}$ can be modeled as a Markov process $p_{\mathcal{T},q'}$. Associated with every state $\mathbf{s}\in\mathcal{T}$ is a distribution $q'_\mathbf{s}\in\mathcal{P}^+(\mathcal{A}\times\mathcal{A})$ which assigns any input/output pair $(a,b)\in\mathcal{A}\times\mathcal{A}$ the probability

$$q'_\mathbf{s}(a,b) = P(X_j = a, Y_j = b|\mathbf{c}_\mathcal{T}(Y_{-\infty}^{j-1}) = \mathbf{s}), \quad \forall j\in\mathbb{N}$$

For convenience, we also denote the input/output transition probabilities encountered upon seeing context $\mathbf{s}\in\mathcal{T}$ by

$$\theta_\mathbf{s}(b|a) = P(Y_1 = b|\mathbf{c}_\mathcal{T}(Y_{-\infty}^0) = \mathbf{s}, X_1 = a).$$

Therefore, we have $q'_\mathbf{s}(a,b) = p_a\,\theta_\mathbf{s}(b|a)$.

The set $\Theta_\mathbf{s} = \{\theta_\mathbf{s}(\cdot|a)\colon a\in\mathcal{A}\}$ is the set of all conditional probabilities associated with state $\mathbf{s}$. Note that $\theta_\mathbf{s}(\cdot|a)\in\mathcal{P}^+(\mathcal{A})$ for all $a\in\mathcal{A}$ and $\mathbf{s}\in\mathcal{T}$. The set $\mathbf{\Theta}_\mathcal{T} = \bigcup_{\mathbf{s}\in\mathcal{T}}\Theta_\mathbf{s}$ is the set of all transition probabilities of channel model and

[1] Irreducible since $q_s\in\mathcal{P}^+(\mathcal{A})$, aperiodic since any state $\mathbf{s}\in\mathcal{T}$ can be reached from itself in either $|\mathbf{s}|$ or $|\mathbf{s}|+1$ steps.

we refer to it as the *channel parameters*. Since the input is a known *i.i.d. Bernoulli* process, estimating $q'(\mathcal{T})$ and $\mathbf{\Theta}_{\mathcal{T}}$ are completely equivalent parameterizations.

As emphasized in the introduction, we do not assume the true channel model is known nor do we assume it is fast mixing . Therefore, by using a sample sequence obtained from the channel, we want (i) to approximate as best as possible, the parameter set $\mathbf{\Theta}_{\mathcal{T}}$ (ii) the stationary probabilities $\mu(\mathbf{s})$ of observing an output string $\mathbf{s} \in \mathcal{T}$, and (iii) estimate or at least obtain heuristics of the information rate of the process.

III. Long Memory and Slow Mixing

There are two distinct difficulties in estimating Markov processes as the ones we are interested in. The first is memory that is too long to handle given the size of the sample at hand. The second issue is that even though the underlying process might be ergodic, the transition probabilities are so small such that the process effectively acts like a non-ergodic process given the sample size available. We illustrate these problems in following simple examples.

Example 1. Let $\mathcal{T} = \mathcal{A}^k$ denote a full tree with depth k and $\mathcal{A} = \{0, 1\}$. Assume that $q_{0^k}(1) = 2\epsilon$ and $q_{10^{k-1}}(1) = 1 - \epsilon$ with $\epsilon > 0$ (where 0^k indicates a string with k consecutive zeros), and let $q_{\mathbf{s}}(1) = \frac{1}{2}$ for all other $\mathbf{s} \in \mathcal{T}$. Let $p_{_{\mathcal{T},q}}$ represent the stationary ergodic Markov process associated with this model. Observe that stationary probability of being in state 0^k is $\frac{1}{2^{k+1}-1}$ while all other states have stationary probability $\frac{2}{2^{k+1}-1}$. Let Y_1^n be a realization of this process with initial state $1^k \preceq Y_{-\infty}^0$. Suppose $k \gg \omega(\log n)$. [2] With high probability we will never find a string of $k-1$ zeros among n samples, and every bit is generated with probability 1/2. Thus with this sample size, with high probability, we cannot distinguish the long-memory process $p_{_{\mathcal{T},q}}$ from an *i.i.d. Bernoulli*$(1/2)$ process. □

We therefore require that dependencies among the transition probabilities of nodes with same parents die down as we look further in the past. This condition will formally be introduced by equation (3) in section IV.

Example 2. Let $\mathcal{A} = \{0, 1\}$ and $\mathcal{T} = \{0, 1\}$ with $q_1(1) = 1 - \epsilon$, and $q_0(1) = \epsilon$. For $\epsilon > 0$, this model represents a stationary ergodic Markov processes with stationary distributions $\mu(1) = \frac{1}{2}, \mu(0) = \frac{1}{2}$. Let $\mathcal{T}' = \{0, 1\}$ with $q'_1(1) = 1 - \epsilon, q'_0(1) = 2\epsilon$. Similarly, for $\epsilon > 0$ this model represents a stationary ergodic Markov processes with stationary distributions $\mu'(1) = \frac{2}{3}, \mu'(0) = \frac{1}{3}$. Suppose we have a length-n sample. In this case, we cannot distinguish between these two models if $\epsilon \ll o(1/n)$, and therefore no estimator can obtain their stationary probabilities either. □

A. Lower Bound on Information Rate

Consider a channel with state tree $\mathcal{T}$ and parameter set $\mathbf{\Theta}_{\mathcal{T}}$. Suppose that $\kappa(\mathcal{T}) = K < \infty$. The information rate for an *i.i.d.* input process with $P(X_i = a) = p_a$ for this channel is

$$R_{\mathcal{T}} \stackrel{\text{def}}{=} \lim_{i \to \infty} \frac{1}{i} I(X_1^i; Y_1^i) = \sum_{\mathbf{s} \in \mathcal{T}} \mu(\mathbf{s}) R_{\mathbf{s}}(\Theta_{\mathbf{s}}) \tag{2}$$

where $R_{\mathbf{s}}(\Theta_{\mathbf{s}})$ is obtained by

$$R_{\mathbf{s}}(\Theta_{\mathbf{s}}) = \sum_{a \in \mathcal{A}} p_a \sum_{b \in \mathcal{A}} \theta_{\mathbf{s}}(b|a) \log \frac{\theta_{\mathbf{s}}(b|a)}{\sum_{a' \in \mathcal{A}} p_{a'} \theta_{\mathbf{s}}(b|a')}.$$

As a remark, note that for fixed input distribution, $R_{\mathbf{s}}$ is a function of $\Theta_{\mathbf{s}} = \{\theta_{\mathbf{s}}(\cdot|a) \colon a \in \mathcal{A}\}$. Furthermore

Lemma 1. $R_{\mathbf{s}}(\Theta_{\mathbf{s}})$ is convex in $\Theta_{\mathbf{s}}$. □

Since the memory is unknown a-priori, a natural approach, known to be consistent, is to use a potentially coarser model with depth k_n. Here k_n increases logarithmically with the sample size n, and reflects [2] well known results on consistent estimation of Markov processes. We show that coarser models formed by properly aggregating states of the original channel are useful in lower bounding information rates of the channel.

Definition 2. Suppose $\tilde{\mathcal{T}} = \mathcal{A}^k$ for some $k \in \mathbb{N}$. The *aggregation* of $p_{_{\mathcal{T},q}}$ at level k, denoted by $p_{_{\tilde{\mathcal{T}},\tilde{q}}}$, is a stationary Markov process with state transition probabilities given by

$$\tilde{q}_{\mathbf{w}}(a) \stackrel{\text{def}}{=} p_{_{\mathcal{T},q}}(a|\mathbf{w}) = \frac{\sum_{\mathbf{v} \in \mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}) q_{\mathbf{v}}(a)}{\sum_{\mathbf{v}' \in \mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}')},$$

for all $\mathbf{w} \in \tilde{\mathcal{T}}$ and $a \in \mathcal{A}$, where μ is the stationary distribution associated with $p_{_{\mathcal{T},q}}$. □

Remark Using Observation 1, wolog, no matter what $\tilde{\mathcal{T}}$ is, we will assume $p_{_{\mathcal{T},q}}$ has states $\mathcal{T}$ such that $\tilde{\mathcal{T}} \preceq \mathcal{T}$. □

Lemma 2. If $p_{_{\tilde{\mathcal{T}},\tilde{q}}}$ aggregates $p_{_{\mathcal{T},q}}$ (with stationary distribution μ), then the stationary distribution of $p_{_{\tilde{\mathcal{T}},\tilde{q}}}$, $\tilde{\mu}$, satisfies for all $\mathbf{w} \in \tilde{\mathcal{T}}$

$$\tilde{\mu}(\mathbf{w}) = \sum_{\mathbf{v} \in \mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}). \qquad \square$$

Example 3. This example illustrates the computations in Definition above. Let $p_{_{\mathcal{T},q}}$ be a Markov process with $\mathcal{T} = \{11, 01, 0\}$ and $q_{11}(1) = \frac{1}{4}, q_{01}(1) = \frac{1}{3}, q_0(1) = \frac{3}{4}$. For this model, we have $\mu(11) = \frac{4}{25}, \mu(01) = \frac{9}{25}$ and $\mu(0) = \frac{12}{25}$. Fig. 1. (b) shows an aggregated process $p_{_{\tilde{\mathcal{T}},\tilde{q}}}$ with $\tilde{\mathcal{T}} = \{1, 0\}$. Note that $\tilde{q}_1(1) = \left(\frac{4}{25}\frac{1}{4} + \frac{9}{25}\frac{1}{3}\right)/\left(\frac{4}{25} + \frac{9}{25}\right) = \frac{4}{13}$ and $\tilde{q}_0(1) = \frac{3}{4}$. Furthermore, we have $\tilde{\mu}(0) = \frac{12}{25}$ and $\tilde{\mu}(1) = \frac{13}{25}$. □

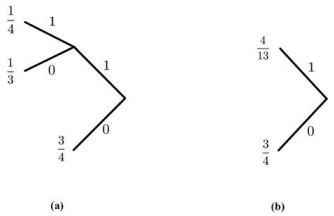

Fig. 1. (a) Markov process in Example 3, (b) Aggregated model at depth 1. From Observation 1, the model on the left can be reparameterized to be a complete tree at any depth ≥ 2. We can hence ask for its aggregation at any depth. Aggregations of the above model on the left at depths ≥ 2 will hence be the model itself.

[2] A function $f_n = \omega(g_n)$ if $\lim_{n \to \infty} f_n / g_n = \infty$.

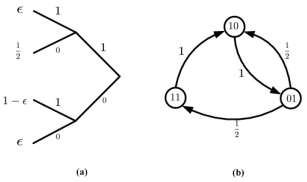

Fig. 2. (a) Markov in Example 4, (b) Same process when $\epsilon = 0$.

Similar to Definition 2, given any input output process for a channel $(\tilde{\mathcal{T}}, \mathbf{\Theta}_{\mathcal{T}})$ we can define an aggregated channel with tree $\tilde{\mathcal{T}} = \mathcal{A}^k$ and parameter set $\tilde{\mathbf{\Theta}}_{\tilde{\mathcal{T}}}$. For all $\mathbf{w} \in \tilde{\mathcal{T}}$ and $a \in \mathcal{A}$, the aggregated model at depth k has parameters

$$\tilde{\theta}_{\mathbf{w}}(b|a) = \frac{\sum_{\mathbf{v}\in\mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v})\theta_{\mathbf{v}}(b|a)}{\sum_{\mathbf{v}'\in\mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}')}, \quad \forall b \in \mathcal{A}.$$

Proposition 1. If $\tilde{\mathcal{T}}$ aggregates $\mathcal{T}$, then $R_{\tilde{\mathcal{T}}} \le R_{\mathcal{T}}$. □

Remark As our results will show, in the slow mixing regime it is not possible to obtain a simple lower bound on the information rate using the data and taking recourse to the Proposition above. Instead, we introduce the *partial* information rate that can be reliably obtained from the data

$$R^p_{\tilde{G}} = \sum_{\mathbf{w}\in\tilde{G}} \frac{\mu(\mathbf{w})}{\mu(\tilde{G})} R_{\mathbf{w}}(\tilde{\Theta}_{\mathbf{w}}),$$

where $\tilde{G} \subseteq \tilde{\mathcal{T}}$ will be a set of *good* states that we show how to identify. The partial information rate is not necessarily a lower bound, but in slow mixing cases it is sometimes the best heuristic possible. □

Notwithstanding the previous remark, we will focus on estimating the aggregated parameters at depth k_n where $k_n = \alpha_n \log n$ for some function $\alpha_n = \mathcal{O}(1)$. However, note that we only have samples from the true channel $\mathbf{\Theta}_{\mathcal{T}}$, not the aggregated channel. It is therefore important to note that our formulation is *not* equivalent to estimating a channel with memory k_n. See also remark after Definition 3.

Example 4. Let $\mathcal{T} = \{11, 01, 10, 00\}$ with $q_{11}(1) = \epsilon$, $q_{01}(1) = \frac{1}{2}$, $q_{10}(1) = 1 - \epsilon$, $q_{00}(1) = \epsilon$. If $\epsilon > 0$, then $p_{\mathcal{T},q}$ is a stationary ergodic Markov process. Let μ denote the stationary distribution of this process. A simple computation shows that $\mu(11) = \frac{1}{7-6\epsilon}$, $\mu(01) = \frac{2-2\epsilon}{7-6\epsilon}$, $\mu(10) = \frac{2-2\epsilon}{7-6\epsilon}$ and $\mu(00) = \frac{2-2\epsilon}{7-6\epsilon}$, and $\mu(1) = \frac{3-2\epsilon}{7-6\epsilon}$ and $\mu(0) = \frac{4-4\epsilon}{7-6\epsilon}$.

Suppose we have a length n sample. If $\epsilon \ll \frac{1}{n}$, then $\mu(1) \approx \frac{3}{7}$ and $\mu(0) \approx \frac{4}{7}$. If the initial state belongs to $\{11, 01, 10\}$, the state 00 will not be visited with high probability in n samples, and it can be seen that the counts of 1 or 0 will not be near the stationary probabilities $\mu(1)$ or $\mu(0)$. For this sample size, the process effectively acts like the irreducible, aperiodic Markov chain in Fig. 2. (b) which in turn is fast mixing. Hence, the stationary probabilities of the chain in Fig. 2. (b), $\frac{\mu(01)}{\mu(1)+\mu(01)}$, $\frac{\mu(10)}{\mu(1)+\mu(01)}$ and $\frac{\mu(11)}{\mu(1)+\mu(01)}$ converge quicker than $\mu(1)$ or $\mu(0)$. This observation guides our formulation of Theorem 4. □

IV. Estimation of Channel Properties

As noted before in Example 1, if the dependencies could be arbitrary in a channel model, we will not estimate the model accurately no matter how large the sample is. Keeping in mind Observation 1, we formalize dependencies dying down by means of a function $d : \mathbb{Z}^+ \to \mathbb{R}^+$ with $\sum_{i=1}^{\infty} d(i) < \infty$. Let $\mathcal{M}_d$ be the set of all channel models $(\mathcal{T}, \mathbf{\Theta}_{\mathcal{T}})$ that satisfy for all $\mathbf{u} \in \mathcal{A}^*$ and all $c, c' \in \mathcal{A}$

$$\left|\frac{\theta_{c\mathbf{u}}(b|a)}{\theta_{c'\mathbf{u}}(b|a)} - 1\right| \le d(|\mathbf{u}|) \tag{3}$$

where $a, b \in \mathcal{A}$ and $\theta_{c\mathbf{u}}(b|a)$ should be interpreted as $P(Y_1 = b|\mathbf{c}_{\mathcal{T}}(Y^0_{-\infty}) = c\mathbf{u}, X_1 = a)$. $\mathcal{M}_d$ has bounded memory iff there exists a finite K such that $d(i) = 0$ for all $i > K$. Note also that $\{d(i)\}_{i\ge1}$ does not control the mixing properties of the channel.

As mentioned in the last section, we will focus on set of the aggregated parameters of a complete tree $\tilde{\mathcal{T}}$, $\Theta_{\tilde{\mathcal{T}}}$, where $\tilde{\mathcal{T}} = \mathcal{A}^{k_n}$ and $k_n = \alpha_n \log n$ for some function $\alpha_n = \mathcal{O}(1)$. But $p_{\tilde{\mathcal{T}},\tilde{q}}$ is still unknown, neither do we have access to samples from it. Therefore, in order to obtain the parameters of the aggregated model, $\tilde{\mathbf{\Theta}}_{\tilde{\mathcal{T}}}$, we use a *naive* estimator—we simply pretend that our sample was in fact from $p_{\tilde{\mathcal{T}},\tilde{q}}$. Equivalently, the estimator pretends that for any output sequence $\mathbf{w} \in \tilde{\mathcal{T}}$, the subsequence of output symbols in the sample that follow $\mathbf{w}$ and associated with the same input letter a is *i.i.d.*.

Throughout this section, we assume that we start with some past $Y^0_{-\infty}$, and we see n samples (X^n_1, Y^n_1) from the channel. All confidence probabilities are conditional probabilities on Y^n_1 given X^n_1 and $Y^0_{-\infty}$. Note that the results hold for all $Y^0_{-\infty}$ (not just a set with probability 1).

Even in the slow mixing case, we want to see if any estimator can be accurate at least partially. In particular, we consider the naive estimator that operates on the assumption that samples are from the aggregated model $p_{\tilde{\mathcal{T}},\tilde{q}}$. There is no reason that the naive estimates should reflect the parameters associated with the true model $p_{\mathcal{T},q}$.

Definition 3. For all sequences (X^n_1, Y^n_1) obtained from the channel model $(\mathcal{T}, \mathbf{\Theta}_{\mathcal{T}})$, let $\tilde{\mathcal{T}} = \mathcal{A}^{k_n}$ with $k_n = \alpha_n \log n$ for some function $\alpha_n = \mathcal{O}(1)$. For $\mathbf{s} \in \tilde{\mathcal{T}}$, let $\mathbf{Y}^a_{\mathbf{s}}$ be the sequence of output symbols that follows the output string $\mathbf{s}$, and correspond to the input symbol a. Hence, the length of $\mathbf{Y}^a_{\mathbf{s}}$ is $N_{\mathbf{s}}(a) = \sum_{j=1}^{n} \mathbb{1}\{\mathbf{c}_{\tilde{\mathcal{T}}}(Y^{j-1}_{-\infty}) = \mathbf{s}, X_j = a\}$ and the number of occurrences of symbol b in $\mathbf{Y}^a_{\mathbf{s}}$ is $N_{\mathbf{s}}(b, a)$, where $N_{\mathbf{s}}(b, a) = \sum_{j=1}^{n} \mathbb{1}\{\mathbf{c}_{\tilde{\mathcal{T}}}(Y^{j-1}_{-\infty}) = \mathbf{s}, Y_j = b, X_j = a\}$. We define the naive estimate of $\tilde{\theta}_{\mathbf{s}}(b|a)$ as

$$\hat{\theta}_{\mathbf{s}}(b|a) \stackrel{\text{def}}{=} \frac{N_{\mathbf{s}}(b, a)}{N_{\mathbf{s}}(a)}.$$

Furthermore, let $N_{\mathbf{s}} = \sum_{a\in\mathcal{A}} N_{\mathbf{s}}(a)$. □

Remark Note that $\mathbf{Y}^a_{\mathbf{s}}$ is *i.i.d.* only if $\mathbf{s} \in \mathcal{T}$, the set of states for the true model. In general, since we do not necessarily know if any of $N_{\mathbf{s}}(b, a)$ are close to their stationary frequencies, there is no obvious reason why $\hat{\theta}_{\mathbf{s}}(b|a)$ shall reflect $\tilde{\theta}_{\mathbf{s}}(b|a)$. □

Somewhat surprisingly, we show that if k_n is large enough, using an argument based on universal compression [13], we show that both the underlying and aggregated parameters will then be close to the naive estimates for frequent states.

Definition 4. Let $\delta_j = \sum_{i \geq j} d(i)$. Note that $\delta_j \to 0$ as $j \to \infty$ and that $-\delta_j \log \delta_j \to 0$ as $\delta_j \to 0$. Given a sample sequence with size n obtained from the channel model $(\mathcal{T}, \mathbf{\Theta}_{\mathcal{T}})$, we define the set of *good* states, denoted by $\tilde{G}$, as

$$\tilde{G} = \left\{ \mathbf{w} \in \tilde{\mathcal{T}} : \forall a \in \mathcal{A}, N_{\mathbf{w}}(a) \geq \max \left\{ n\delta_{k_n} \log \frac{1}{\delta_{k_n}}, 2^{k_n+1} \log^2 n \right\} \right\}.$$

Remark Note that a state is good if the count of the state is $\geq n^{\alpha_n} \log^2 n = 2^{k_n} \log^2 n$. Therefore, if $2^{k_n} \log^2 n \geq n$, or equivalently $k_n \geq \log n - 2 \log \log n$, no state will be good and the Theorem below becomes vacuously true. This is not a fundamental weakness in this line of argument—it is known that k_n has to scale logarithmically with n for proper estimation to hold. □

Theorem 3. Let $(\mathcal{T}, \mathbf{\Theta}_{\mathcal{T}})$ be an unknown channel in $\mathcal{M}_d$. If $k_n = \alpha_n \log n$, then with probability (under the true model $(\mathcal{T}, \mathbf{\Theta}_{\mathcal{T}})$ and conditioned on $Y_{-\infty}^0$) $\geq 1 - \frac{1}{2^{|\mathcal{A}|^{k_n+1} \log n}}$, for all $a \in \mathcal{A}$ and $\mathbf{w} \in \tilde{G}$ simultaneously

$$||\tilde{\theta}_{\mathbf{w}}(\cdot|a) - \hat{\theta}_{\mathbf{w}}(\cdot|a)||_1 \leq 2\sqrt{\frac{\ln 2}{\log n} - \frac{\ln 2}{\log \delta_{k_n}}}. \qquad \square$$

Remark Since we do not assume the source has mixed, the above theorem does *not* imply that the parameters are accurate for contexts shorter than k_n. While perhaps counter-intuitive at first glance, note the above result does not depend on empirical counts being near stationary probabilities. □
When the dependencies among strings die down exponentially, we can strengthen Theorem 3 to get convergence rate polynomial in n as in [13].

Note that the aggregated parameters associated with any $\mathbf{w} \in \tilde{G}$ can be approximately estimated from the sample (obtained from the true channel) while the rest may not be accurate. But from Example 2, we know that the stationary probabilities may be a very sensitive function of the parameters associated with states. How do we tell, therefore, if we can trust our naive counts of states?

To find deviation bounds for stationary distribution of good states, we construct a new process $\{Z_m\}_{m \geq 1}$, $Z_m \in \mathcal{T}$ from the process $\{Y_i\}_{i \geq 1}$. If Y_{i_m} is the $(m+1)^{th}$ symbol in the sequence $\{Y_i\}_{i \geq 1}$ such that $\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{i_m}) \in \tilde{G}$, then $Z_m = \mathbf{c}_{\mathcal{T}}(Y_{-\infty}^{i_m})$. The strong Markov property allows us to characterize $\{Z_m\}_{m \geq 1}$ as a Markov process with transitions that are lower bounded by those transitions of the process $\{Y_i\}_{i \geq 1}$ that can be well estimated by the Theorem above. Note that even if the original process is aperiodic, it is quite possible that $\{Z_m\}_{m \geq 1}$ be aperiodic.

For any (good) state $\mathbf{w}$, let $G_{\mathbf{w}} \subset \mathcal{A}$ be the set of letters that take $\mathbf{w}$ to another good state, $G_{\mathbf{w}} = \{b \in \mathcal{A} : \mathbf{c}_{\tilde{\mathcal{T}}}(\mathbf{w}b) \in \tilde{G}\}$. Our confidence in the empirical counts of good states matching their (aggregated) stationary probabilities follows from a coupling argument [13], and depends on the following parameter

$$\eta_{\tilde{G}} = \min_{\mathbf{u}, \mathbf{v} \in \tilde{G}} \sum_{b \in G_{\mathbf{u}} \cap G_{\mathbf{v}}} \min \{\tilde{q}_{\mathbf{u}}(b), \tilde{q}_{\mathbf{v}}(b)\}, \qquad (4)$$

where $\tilde{q}_{\mathbf{u}}(b) \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} p_a \tilde{\theta}_{\mathbf{u}}(b|a)$.

The counts of various $\mathbf{w} \in \tilde{G}$ now concentrates as shown in the following Theorem, and how good the concentration is can be estimated as a function of $\eta_{\tilde{G}}$ (and δ_{k_n}) and the total count of all states in $\tilde{G}$ as below. Now $\tilde{G}$ as well as $\eta_{\tilde{G}}$ are well estimated from the sample using Theorem 3—thus we can look at the data to interpret the empirical counts of various substrings of the data.

Let $\Delta_j = \sum_{i \geq j} \delta_i$. For the following theorem, we require $\{\delta_i\}_{i \geq 1}$ to be summable—stronger condition than Theorem 3. We assume that $\delta_i \leq \frac{1}{i}$.

Theorem 4. If $\{Z_m\}_{m=1}^{\infty}$ is aperiodic, then for any $t > 0$, $Y_{-\infty}^0 \in \mathcal{A}^{\infty}$ and $\forall \mathbf{w} \in G$

$$p_{\mathcal{T},q}(|N_{\mathbf{w}} - \tilde{n}\frac{\mu(\mathbf{w})}{\mu(\tilde{G})}| \geq t | Y_{-\infty}^0) \leq 2 \exp\left(-\frac{t^2}{2\tilde{n}} \left(\frac{\eta_{\tilde{G}}^{k_n}(1-\Delta_{k_n})}{4\ell_n + \eta_{\tilde{G}}^{k_n}(1-\Delta_{k_n})} \right)^2 \right)$$

where ℓ_n is the smallest integer such that $\Delta_{\ell_n} \leq \frac{1}{n}$, $\tilde{n}$ is the total count of good states in the sample and μ is the stationary distribution of $p_{\mathcal{T},q}$. □

REFERENCES

[1] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*. Americam Mathematical Society, 2009.

[2] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via bic and mdl," *IEEE Transactions on Information theory*, vol. 52, no. 3, Mar 2006.

[3] A. Garivier, "Consistency of the unlimited BIC context tree estimator," *IEEE Transactions on Information theory*, vol. 52, no. 10, pp. 4630–4635, Sep 2006.

[4] A. Garivier and F. Leonardi, "Context tree selection: A unifying view," *Stochastic Processes and their Applications*, vol. 121, no. 11, pp. 2488–2506, Nov 2011.

[5] A. Galves, V. Maume-Deschamps, and B. Schmitt, "Exponential inequalities for VLMC empirical trees," *ESAIM: Probability and Statistics*, vol. 12, pp. 219–229, Jan 2008.

[6] P. Bühlmann and A. Wyner, "Variable length Markov chains," *Annals of Statistics*, vol. 27, no. 2, pp. 480–583, 1999.

[7] J. Rissanen, "A universal data compression system," *IEEE Transactions on Information theory*, vol. 29, no. 5, pp. 656–664, Sep 1983.

[8] I. Csiszár and Z. Talata, "On rate of convergence of statistical estimation of stationary ergodic processes," *IEEE Transactions on Information theory*, vol. 56, no. 8, pp. 3637–3641, Aug 2010.

[9] I. Csiszár, "Large-scale typicality of Markov sample paths and consistency of MDL order estimators," *IEEE Transactions on Information theory*, vol. 48, no. 6, pp. 1616–1628, Jun 2002.

[10] I. Csiszár and P. C. Shields, "The consistency of the BIC Markov order estimator," *Annals of Statistics*, vol. 28, pp. 1601–1619, 2000.

[11] D. J. K. Farzan, "Coding schemes for chip-to-chip interconnect applications," *IEEE Transansactions on Very Large Scale Integration Systems*, vol. 14, no. 4, pp. 393–406, Apr 2006.

[12] D. J. Aldous, "Random walks on finite groups and rapidly mixing Markov chains," in *Séminaire de Probabilités XVII - 1981/82, Springer Lecture Notes in Mathematics 986*, 1983.

[13] M. Asadi, R. Paravi, and N. P. Santhanam, "Estimation in slow mixing, long memory channels," available on http://arxiv.org/abs/1301.6798.

[14] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: basic properties." *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[15] W. Feller, *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons; 2nd Edition, 1957, vol. 1.