The Annals of Statistics
2017, Vol. 45, No. 1, 355–385
DOI: 10.1214/16-AOS1455

APPROXIMATE GROUP CONTEXT TREE

BY ALEXANDRE BELLONI AND ROBERTO I. OLIVEIRA[1]

The Fuqua School of Business
Duke University and IMPA

We study a variable length Markov chain model associated with a group of stationary processes that share the same context tree but each process has potentially different conditional probabilities. We propose a new model selection and estimation method which is computationally efficient. We develop oracle and adaptivity inequalities, as well as model selection properties, that hold under continuity of the transition probabilities and polynomial β-mixing. In particular, model misspecification is allowed.

These results are applied to interesting families of processes. For Markov processes, we obtain uniform rate of convergence for the estimation error of transition probabilities as well as perfect model selection results. For chains of infinite order with complete connections, we obtain explicit uniform rates of convergence on the estimation of conditional probabilities, which have an explicit dependence on the processes' continuity rates. Similar guarantees are also derived for renewal processes.

Our results are shown to be applicable to discrete stochastic dynamic programming problems and to dynamic discrete choice models. We also apply our estimator to a linguistic study, based on recent work by Galves et al. [*Ann. Appl. Stat.* **6** (2012) 186–209], of the rhythmic differences between Brazilian and European Portuguese.

1. Introduction. In this paper, we are interested in applying *context tree models*, also known *variable length Markov chains* (VLMCs), to the estimation of transition probabilities and dependence structures in discrete-alphabet stochastic processes. Context tree models describe processes where each infinite "past" has a finite suffix—the *context*—that suffices to determine the transition probabilities. As such, they are generalizations of finite-order Markov chains, for which contexts exist and are of fixed length. Context tree processes first appeared in Rissanen's seminal paper [24], where two appealing traits were noted.

- *Parsimony*: a Markov chain model must have an order parameter that is large enough to distinguish any two pasts with different transition probabilities. By contrast, by using different context lengths for different pasts, one may need

Received February 2015; revised December 2015.
[1]Supported by projects *Universal* and *Produtividade em Pesquisa* from CNPq, Brazil.
MSC2010 subject classifications. Primary 62M05, 62M09, 62G05; secondary 62P20, 60J10.
Key words and phrases. Categorical time series, group context tree, dynamic discrete choice models, dynamic programming, model selection, VLMC.

less parameters to specify the model. (Incidentally, this motivates the VLMC terminology.)

- *Computationally efficient estimation*: the set of context has a natural suffix tree structure, known as the *context tree*. The fact that this is a tree allows for efficiency search over an exponentially large class of models. Rissanen's original Context algorithm for estimating the context tree relied strongly on this.

Both traits have continued to play a role over the years as a growing number of papers on context tree models appeared in Statistics [10, 11, 16], Information Theory [18, 27, 28], Bioinformatics [3] and Linguistics [17]. In this last paper, *interpretability* of context trees has also played a role, which adds to their interest as practical tools.

In this paper, we consider context tree model selection and estimation for a *group* of $\mathrm{L} \geq 1$ stationary processes over a discrete alphabet. These stationary processes have the same context tree but possibly different conditional probability distributions. We refer to this model as *group context tree* alluding to the recent literature on group lasso [21, 22, 29]. As in the case of group lasso, by combining different processes with similar dependence structure we hope to improve the overall estimation. In addition, the model we consider also allows for processes which are only approximately described by a finite context tree, hence the name *approximate group context tree* (AGCT) model.

Although this group context tree setting is new, our estimator and the results we obtain are related to several papers that considered a single stationary process ($\mathrm{L} = 1$), which we outline briefly. Bühlmann and Wyner [11] proved properties of the Context estimator allowing the model to grow with the sample size. They also studied a bootstrap scheme based on fitted VLMCs. Ferrari and Wyner [16] consider processes with infinite dependence for which there exist "good" context tree approximations. They established new results on a sieve methodology based on an adaptation of the Context algorithm. The BIC Context Tree algorithm and its consistency properties have been considered in [14, 18] and [26]. Redundancy rates were studied by [13] and [18]. Several other works contributed to this literature in various directions; see [9, 10, 19, 27] and the references therein.

In Section 2, we propose an estimator for model selection and estimation of conditional probabilities based on context tree models, which *does not assume a true VLMC model*. As in Rissanen's original estimator, we first build a full suffix tree for the observed sample, then prune the tree by removing "statistically insignificant" nodes. In addition to considering a group of processes, the proposed estimator also differs on how we define insignificance. We use a procedure reminiscent of Lepskii's adaptation method [20]. For each suffix, we compute from the sample an (approximate) confidence radius for its vector of transition probability estimates (one for each process). We then recursively prune any leaf node w whose descendants w' in the full sample suffix tree (i.e., the tree prior to pruning) are all "compatible" with the parent of w, in the sense that the corresponding confidence

regions intersect. By a judicious choice of confidence radii, this procedure automatically balances the variance coming from the random sampling with the bias incurred by the truncation mechanism.

Section 3 details the assumptions we impose on processes, most notably *continuity of transition probabilities* (deeper truncation implies arbitrarily good approximation). Based on this, finite sample results on adaptivity and model selection properties are presented in Section 4. In that same section, we present stronger results, including oracle inequalities, that require an added assumption of *polynomial β-mixing*. Previous work in the area imposed assumptions that implied a true finite VLMC model, exponential mixing properties and/or nonnullness (positivity) of the transition probabilities, which we manage to avoid here. Moreover, our oracle inequality for the AGCT estimator (Corollary 1) seems to be the first result of its kind for context tree estimation, even in the single process case.

In Section 5, we present three classes of examples where our general results can be applied. For parametric models (i.e., actual finite-order Markov chains), we derive uniform rate of convergence to transition probabilities, as well as perfect model selection, under weaker assumptions than [11] (which only covered the single process case). For chains of infinite order with complete connections, we obtain explicit uniform rates of convergence on the estimation of conditional probabilities, which have an explicit dependence on the processes' continuity rates. We also derive explicit uniform rates of convergence for certain renewal processes. In most cases, we show that the group context tree model can lead to improvements on the estimation when compared to the single-process case.

Group context tree models are used in Section 6 to estimate dynamic marginal effects in dynamic choice models [1, 7], and to estimate the value function in discrete stochastic dynamic programming problems [5, 15, 23, 25]. In these applications, the objects of main interest are functionals of the conditional probabilities. We derive uniform bounds on the rate of convergence for the estimates that hold uniformity over all possible contexts and account for model selection mistakes. Furthermore, in Section 7 we revisit a study by Galves et al. [17], and apply the AGCT model to understand the difference between the rhythmic features in European and Brazilian Portuguese. A key point is that the AGCT framework allows for the processes to have different transition probabilities.

Section 8 discusses variations of the estimator and comparisons, and a final section adds some further thoughts. Proofs are mostly contained in two Appendices. Simulations and some auxiliary theoretical results are provided in the supplementary material [4].

1.1. *Notation*. Let A denote a finite set (called alphabet), and the set of probability distributions over A will be denoted by Δ^A. We use A_{-k}^{-1} to denote all A-valued sequences with length k, and $A^* = A_{-\infty}^{-1} \cup (\bigcup_{k=0}^{\infty} A_{-k}^{-1})$. The length of a string w is denoted by $|w|$ and, for each $1 \le k \le |w|$, w_{-k}^{-1} is the suffix

of w with length k. We also let $w_0^{-1} = e$, the empty string. A subset $\tilde{T} \subset A^*$ is a *tree* if the empty string $e \in \tilde{T}$ and for all $w = w_{-|w|} \cdots w_{-1} \in \tilde{T} \setminus \{e\}$ the string $w_{-k}^{-1} = w_{-k} \cdots w_{-1} \in \tilde{T}$ for any $k \leq |w|$. The *parent* of w is denoted by $\mathrm{par}(w) = w_{-|w|+1} \cdots w_{-1}$. An element of a tree $\tilde{T}$ that is not the parent of any other element in $\tilde{T}$ is said to be a *leaf* of $\tilde{T}$. For $w, w' \in A^*$, we write $w \preceq w'$ if w is a suffix of w'.

We associate with each tree $\tilde{T}$ and each $x = \cdots x_{-3}x_{-2}x_{-1} \in A_{-\infty}^{-1}$ a suffix $\tilde{T}(x)$ of x with the following rule:

- If $x_{-k}^{-1} \in \tilde{T}$ for all $k \in \mathbb{N}$, then $\tilde{T}(x) = x$;
- Otherwise, take the largest $k \in \mathbb{N}$ with $x_{-k}^{-1} \in \tilde{T}$ and set $\tilde{T}(x) = x_{-k}^{-1}$. (Note that this is the empty string if $k = 0$.)

The strings of the form $\tilde{T}(x)$ where x ranges over $A_{-\infty}^{-1}$ will be called the *terminal nodes* of $\tilde{T}$. Notice that all terminal nodes are either leaves or infinite strings. For two sequences a_n, b_n we denote $a_n \lesssim b_n$ if $a_n = O(b_n)$. The indicator function of an event E is denoted by 1_E, and for $q \geq 1$ the $\|\cdot\|_{\mathrm{L},q}$-norm of a vector $v \in \mathbb{R}^{\mathrm{L}}$ is defined as

$$\|v\|_{\mathrm{L},q} = \left(\frac{1}{\mathrm{L}} \sum_{\ell=1}^{\mathrm{L}} |v_\ell|^q \right)^{1/q}. \tag{1.1}$$

2. Setting for group context trees. A pair $(\tilde{T}, \tilde{p})$ will correspond to a tree $\tilde{T}$ and a mapping $\tilde{p}$ that assigns to each terminal node v of $\tilde{T}$ a probability distribution $\tilde{p}(\cdot|v)$ over a finite alphabet A. A stationary ergodic process $X \equiv (X_k)_{k\in\mathbb{Z}}$ will be said to be *compatible with* $(\tilde{T}, \tilde{p})$ if

$$\mathbb{P}\big(X_0 = a | X_{-\infty}^{-1}\big) = \tilde{p}\big(a|\tilde{T}\big(X_{-\infty}^{-1}\big)\big) \qquad \text{almost surely.}$$

On a group context tree model, we have a family $X = (X(\ell))_{\ell=1}^{\mathrm{L}}$ of L independent and stationary processes

$$X(\ell) \equiv \big(X_k(\ell)\big)_{k\in\mathbb{Z}} \qquad (1 \leq \ell \leq \mathrm{L}),$$

a single context tree T^*, and (possibly distinct) probability distributions p_ℓ, $\ell = 1, \ldots, \mathrm{L}$, such that the ℓth process is compatible with (T^*, p_ℓ) for $\ell = 1, \ldots, \mathrm{L}$. Note that T^* is possibly infinite so that this is not a restriction/assumption on the model. Moreover, if the ℓth process is compatible with a context tree T^ℓ, we have $T^* = \bigcup_{\ell=1}^{\mathrm{L}} T^\ell$, and we may redefine p_ℓ correspondingly.

To quantify the approximation error, we use a metric $d_\ell : \Delta^A \times \Delta^A \to [0, 1]$ for each process, $\ell = 1, \ldots, \mathrm{L}$ and write the associated L-vector $d(p, q) = (d_1(p_1, q_1), \ldots, d_{\mathrm{L}}(p_{\mathrm{L}}, q_{\mathrm{L}}))'$. We will aggregate the approximation errors across processes through $\|d(p, q)\|_{\mathrm{L},F}$ where $\|\cdot\|_{\mathrm{L},F}$ is the norm defined in (1.1). For

simplicity, we consider all metrics d_ℓ to be equal and of a certain specific kind. Namely, there exists a collection $\mathcal{S}$ of subsets of A such that

$$d_\ell(p_\ell, q_\ell) = \sup_{S\in\mathcal{S}} \big| p_\ell(S) - q_\ell(S) \big|, \qquad \ell = 1, \dots, \mathrm{L}. \tag{2.1}$$

Our main interests are in the ℓ_1 metric, where $\mathcal{S} = 2^A$ consists of all subsets of A, and the ℓ_∞ metric, where $\mathcal{S}$ consists of all singletons of A.

2.1. *The AGCT estimator.* In this section, we discuss the model selection method which leads to the estimation of the conditional probabilities from a sample of L processes. For each $\ell = 1, \dots, \mathrm{L}$, our sample consists of a string of size n with symbols from A denoted as $X_1^n(\ell) \equiv (X_1(\ell), \dots, X_n(\ell))$. For a string $w \in A^*$, we let $N_{k,\ell}(w)$ denote the number of occurrences of w in $X_1^k(\ell)$.[2] (For notational convenience, we assume that the length n of the sample of each process is the same but the analysis does not rely on that.)

The algorithm proceeds in three steps: Initialization, Identification of Removable Nodes and Pruning. Next, we describe in detail the procedure. In what follows, we let E_n be the suffix tree that contains every string $w \in A^*$ which appears in all L data sequences of the sample X_1^{n-1}, namely

$$E_n = \Big\{ w \in A^* : \min_{\ell=1,\dots,\mathrm{L}} N_{n-1,\ell}(w) > 0 \Big\}. \tag{2.2}$$

Step 1: *Initialization*. For each $w \in E_n$, we specify a conditional probability estimate and a confidence radius:

$$\hat{p}_{n,\ell}(a|w) \equiv \frac{N_{n,\ell}(wa)}{N_{n-1,\ell}(w)} \qquad \text{for } a \in A, \ell = 1, \dots, \mathrm{L}, \quad \text{and} \quad \hat{\mathsf{cr}}_\ell(w). \tag{2.3}$$

The estimator $\hat{p}_{n,\ell}(a|w)$ is a nonparametric estimate for the transition probability $p_\ell(a|w)$. The confidence radius $\hat{\mathsf{cr}}_\ell(w)$, to be specified in Section 2.2 below, depends on the choice of $\digamma$. With high probability, it is essentially an upper bound for the distance between $p(\cdot|w)$ and $\hat{p}_n(\cdot|w)$, up to a bias factor that comes from truncating the past of the process at w (this is related to the *continuity rates*, cf. Assumption 2 below).

Step 2: *Identifying removable nodes*. For a fixed constant $c > 1$, define for each $w \in E_n$:

$$\mathsf{CanRmv}(w) \equiv \begin{cases} 1, \text{ if for all } w', w'' \in E_n \text{ with } w \preceq w', \mathrm{par}(w) \preceq w'', \\ \quad \big\| d\big(\hat{p}_n(\cdot|w'), \hat{p}_n(\cdot|w'')\big) \big\|_{\mathrm{L},\digamma} \le c \big\| \hat{\mathsf{cr}}(w') \big\|_{\mathrm{L},r} + c \big\| \hat{\mathsf{cr}}(w'') \big\|_{\mathrm{L},r}; \\ 0, \text{ otherwise.} \end{cases} \tag{2.4}$$

[2] Formally defined for $k \ge |w| + 1$, so that $N_{k,\ell}(w)$ denotes the number of indices i, $|w| \le i \le k$, with $X_{i-|w|}^i(\ell) = w$.

Intuitively, $\mathsf{CanRmv}(w) = 1$ means that we can remove w, which happens if and only if, for any two nodes $w \preceq w'$, $\mathrm{par}(w) \preceq w''$, the distance between the corresponding transition probability estimates is smaller than the sum of the noise levels at the nodes. The slack factor $c > 1$ allows us to keep a check on the bias that might be incurred by removing w. Our analysis in the Appendix shows that using $c > 1$ implies that, with high probability, this bias will not be much larger than the noise.[3] This is similar, for example, to the slack parameter used in [6], and we recommend $c = 1.01$ in practice.

Step 3: *Pruning*. Let $\hat{T}_n \leftarrow E_n$. Prune any leaf of $\hat{T}_n$ with $\mathsf{CanRmv}(w) = 1$. Repeat until all leaves of $\hat{T}_n$ have $\mathsf{CanRmv}(w) = 0$. Return $(\hat{P}_n, \hat{T}_n)$ where

$$\hat{P}_n(a|x) \equiv \hat{p}_n\big(a|\hat{T}_n(x)\big)$$

for all $x \in A_{-\infty}^{-1}$ and $a \in A$.

This last step keeps the smallest subtree of E_n containing all nodes that *cannot* be removed [i.e., for all $w \in \hat{T}_n$ we have $\mathsf{CanRmv}(w) = 0$]. For completeness, we provide detail algorithm in Figure 1 in Section 4 of the supplementary material [4]. The context tree $\hat{T}_n$ is our selected model, and the transition probability estimate $\hat{P}_n$ is compatible with it. We will show that pruning typically removes high-noise nodes, and the bias incurred by pruning is kept manageable by the test in CanRmv.

2.2. *Data-driven choices of confidence radii*. The performance of our algorithm is heavily dependent on choices of confidence radii $\hat{\mathsf{cr}}_\ell(w)$. As noted above, we will choose those so as to bound from above the deviations $\|d(\hat{p}_n(\cdot|w), p(\cdot|w))\|_{\mathrm{L},F}$ up to an extra error term depending on the continuity rates. There is an important tradeoff between a large confidence radius that introduces a large bias and small confidence radius that do not properly account for the noise in the data. In this section, we present choices that achieve good balance between these factors. These choices ultimately derive from the self-normalized martingale inequalities that we present in the Appendix and supplementary material [4].

DEFINITION 1 (First choice of confidence radius). Let $1-\delta, \delta \in (0,1)$ be our desired confidence level. For $w \in E_n$, $\ell = 1, \ldots, \mathrm{L}$, let

$$\hat{\mathsf{cr}}_\ell(w) \equiv \sqrt{\frac{4}{N_{n-1,\ell}(w)}\left(2\ln\big(2+\log_2 N_{n-1,\ell}(w)\big) + \ln\left(\frac{n^2\mathrm{L}|\mathcal{S}|}{\delta}\right)\right)}.$$

The choice above satisfies $\hat{\mathsf{cr}}_\ell(w) \sim \sqrt{\log(n\mathrm{L}/\delta)/N_{n-1,\ell}(w)}$. This choice exhibits the same behavior as in the case of a single group ($\mathrm{L} = 1$) provided $\log \mathrm{L} \lesssim \log n$, which encompasses most cases of interest. The choice in Definition 1 is desired when we want our estimates of the transition probabilities to be

[3] See the proof of Lemma 2.

uniformly good approximations. The next proposal for confidence radius is appropriate when the number of processes is large and we want our estimates to be good on average.

DEFINITION 2 (Second choice of confidence radius). Let $1-\delta$, $\delta \in (0,1)$, be the desired confidence level. Assume the condition

$$\mathrm{L} \geq 6\ln\left(\frac{n^2}{\delta}\right) \tag{2.5}$$

and for $w \in E_n$, $\ell = 1, \ldots, \mathrm{L}$, let

$$\hat{\mathsf{cr}}_\ell(w) \equiv \sqrt{\frac{4}{N_{n-1,\ell}(w)}\left(2\ln\bigl(2+\log_2 N_{n-1,\ell}(w)\bigr) + \ln|\mathcal{S}| + 1 + \sqrt{\frac{6\ln(\frac{n^2}{\delta})}{\mathrm{L}}}\right)}.$$

In this case, because of (2.5), the rate of $\hat{\mathsf{cr}}_\ell(w)$ is $\sqrt{\log\log n / N_{n-1,\ell}(w)}$ improving upon the single-process case. This is remarkably close to the error in the estimation of probabilities if the model was known in advance.

3. Assumptions. In this section, we state the main assumptions on the processes $(X(\ell))_{\ell=1}^{\mathrm{L}}$ for our main results. For clarity, we decided to use relatively transparent hypotheses, but slightly more general assumptions can be imposed with very few changes.

3.1. *Basic distributional assumptions*. We start with the simplest assumptions that allow for effective use of the group structure, in that we consider the same "prefixes" for all processes. To make this precise, we define the *support* supp_ℓ of process $X(\ell)$ as the set

$$\mathrm{supp}_\ell \equiv \bigl\{x_{-\infty}^{-1} \in A_{-\infty}^{-1} : \forall k \in \mathbb{N}, \mathbb{P}\bigl(X_{-k}^{-1}(\ell) = x_{-k}^{-1}\bigr) > 0\bigr\},$$

and formally state our condition.

ASSUMPTION 1 (Framework). We have L processes

$$X(\ell) = \bigl(X_k(\ell)\bigr)_{k\in\mathbb{Z}}, \qquad 1 \leq \ell \leq \mathrm{L}$$

taking values in the same discrete alphabet A which are independent and stationary. All processes have the same (potentially infinite) context tree T^* and (potentially different) transition probabilities $p_1, \ldots, p_{\mathrm{L}}$. The sets supp_ℓ, $1 \leq \ell \leq \mathrm{L}$, are all equal. We denote by $\mathrm{supp} \equiv \mathrm{supp}_1$. We observe $\{X_1^n(\ell)\}_{\ell=1}^{\mathrm{L}}$, samples of length $n \geq 9$ of the stochastic processes $\{X(\ell)\}_{\ell=1}^{\mathrm{L}}$.

3.2. *Continuity rates and mixing.* The uniform control we aim for essentially requires that truncating the past of the process at some past time $-k$, $k \gg 1$, is not too hurtful for the transition probabilities.

ASSUMPTION 2 (Continuity). The processes $X(\ell)$, $1 \le \ell \le \mathrm{L}$, are *continuous*. That is, for each ℓ, there exists a version of the conditional probabilities p_ℓ of the $X(\ell)$ process such that the quantities

$$\gamma_\ell(x_{-k}^{-1}) \equiv \sup_{y,z \in A_{-\infty}^{-1} : y_{-k}^{-1} = z_{-k}^{-1} = x_{-k}^{-1}} d_\ell\big(p_\ell(\cdot|y),\, p_\ell(\cdot|z)\big)$$

converge to 0 as $k \to +\infty$, for all $x_{-\infty}^{-1}$, where d_ℓ is a metric as in (2.1).

The numbers $\gamma_\ell(\cdot)$ are the continuity rates of process ℓ. A compactness argument implies that their convergence to 0 is *uniform* in $x \in A_{-\infty}^{-1}$. However, our estimator will *adapt* to the continuity rates, meaning that it will tend to do better on pasts that are "more continuous."

4. Finite sample analysis. In this section, we derive our main theoretical results on the performance of the estimates proposed in Section 2.

4.1. *Main results*: *Adaptivity and an oracle inequality.* We can now state our main result.

THEOREM 1 (Main theorem; proven in Appendix A). *Under Assumptions 1 and 2, let $\hat{T}_n$ and $\hat{P}_n$ denote the tree and transition probabilities output by the AGCT algorithm with $\delta \in (0,1)$, $c > 1$ and one of the options below*:

- General case. *We use any $F \in [1, \infty]$, take $r = F$ and use the confidence radii as in Definition 1.*
- Many processes. *In this case, we assume condition (2.5) in Definition 2, take $F = 1$, $r = 2$ and use the confidence radii in that definition.*

Then the following facts hold simultaneously with probability at least $1 - \delta$:

1. *The estimated tree is contained in the correct tree*: $\hat{T}_n \subset T^*$.
2. *Uniformly over $x \in$ supp, we have*

$$\big\| d\big(p(\cdot|x), \hat{P}_n(\cdot|x)\big) \big\|_{\mathrm{L},F} \le \inf_T \frac{2c+2}{c-1} \big\| \gamma\big(T(x)\big) \big\|_{\mathrm{L},F} + (1+2c) \big\| \hat{\mathsf{cr}}\big(T(x)\big) \big\|_{\mathrm{L},r}.$$

Theorem 1 contains two assertions that hold with high probability. First, the AGCT estimator does not give a bigger tree than necessary: this is advantageous when there is a true, finite VLMC model with a small T^*. However, note that, in general, T^* might contain some infinite paths.

Second, Theorem 1 shows that our estimator adapts to the continuity rates of the process in a very strong, pastwise sense. The transition probabilities for more frequent pasts are better approximated because the confidence radii $\hat{\mathsf{cr}}_\ell(T(x))$ decrease when the $N_{n-1,\ell}(T(x))$ increase. This is enough to imply the almost sure converge of the AGCT probability estimates to the transition probabilities for continuous, ergodic processes, when the sample size n increases and the values of $\delta = \delta^{(n)}$ chosen are summable.

An added feature is that, under (2.5), we may use the second choice of confidence radii in Definition 2 (with $F = 1$, $r = 2$) and obtain faster rate of convergence by a $\sqrt{\log n / \log\log n}$ factor relative to the choice in Definition 1. This is indeed the case for some processes studied in more detail in the supplementary material [4].

REMARK 1 (Generality of adaptivity). The result in Theorem 1 holds for any stationary process. The generality of Theorem 1 is achieved through the use of self-normalized martingale inequalities derived in the supplementary material [4]. Those inequalities are used to establish the validity of the data-driven choice of the confidence radius. However, the rates of convergence depend on sample realization through the confidence radius. In order to derive explicit rates of convergence, it is necessary to control how fast the L processes lose memory; see Section 4.2.

4.2. *Main results for β-mixing processes.* In this section, we assume the processes satisfy a polynomial β-mixing condition, which is known to hold for a wide class of processes. (This property can sometimes be derived from the continuity rates; see Section 5.) Recall that a process $X_{-\infty}^{+\infty}$ with values in a finite alphabet A is said to be β-mixing (or absolutely regular) if there exists a function $\beta : \mathbb{N} \to [0, 1]$ with $\lim_{b\in\mathbb{N}, b\to\infty} \beta(b) = 0$ and $\forall k \in \mathbb{Z}, s \in \mathbb{N}$:

$$\beta(b) \geq \mathbb{E}\Big[\sup_{E\subset A^s} \big|\mathbb{P}\big(X_{k+b}^{k+b+s-1} \in E | X_{\infty}^{k}\big) - \mathbb{P}\big(X_{k+b}^{k+b+s-1} \in E\big)\big|\Big].$$

The function $\beta(\cdot)$ is called a (β-)mixing rate function for $X_{-\infty}^{+\infty}$. We assume the following.

ASSUMPTION 3 (Polynomial β-mixing). The L processes $X(1), \ldots, X(\mathrm{L})$ are all polynomially β-mixing with common rate function $\beta(b) \equiv \Gamma b^{-\theta}$ ($b \in \mathbb{N}$), where $\Gamma, \theta > 0$.

This extra assumption will allow us to control how "typical" context trees behave as estimators, which in turn allows us to establish guarantees for the proposed AGCT estimator. To characterize the set of typical trees, recall that under Assumption 1 the processes $X(1), \ldots, X(\mathrm{L})$ have the same support supp, and we define

$$\pi_\ell(w) \equiv \mathbb{P}\big(X_{-|w|}^{-1}(\ell) = w\big) \qquad (w \in \mathrm{supp}, 1 \leq \ell \leq \mathrm{L}). \tag{4.1}$$

For a finite tree T, define π_T as the minimum stationary probability of a leaf node,

$$\pi_T := \min\{\pi_\ell(w) : 1 \le \ell \le \mathrm{L}, w \in \text{supp is a leaf of } T\},$$

and let h_T denote the height of T,

$$h_T := \max\{|w| : w \in \text{supp is a leaf of } T\}.$$

DEFINITION 3 (Typical trees). For (h, π_*), define the set of *typical* trees $\mathcal{T}(h, \pi_*)$ as the set of all finite trees T satisfying $\pi_T \ge \pi_*$ and $h_T \le h$.

Define also the population analogues of confidence radii:

$$\overline{\mathsf{cr}}_\ell(w) \equiv \begin{cases} \sqrt{\dfrac{8}{\pi_\ell(w)n}}\sqrt{2\ln\bigl(2+\log_2\{\pi_\ell(w)n/2\}\bigr)+\ln\Bigl(\dfrac{|\mathcal{S}|\mathrm{L}n^2}{\delta}\Bigr)}, & \text{or} \\ \sqrt{\dfrac{8}{\pi_\ell(w)n}}\sqrt{2\ln\bigl(2+\log_2\{\pi_\ell(w)n/2\}\bigr)+\ln|\mathcal{S}|+1+\sqrt{\dfrac{6}{\mathrm{L}}\ln\Bigl(\dfrac{n^2}{\delta}\Bigr)}}, \end{cases}$$

where $N_{n-1,\ell}(w)$ is replaced by $\pi_\ell(w)n/2$ in $\hat{\mathsf{cr}}_\ell(w)$.

The next result exploits the β-mixing condition to provide finite sample bounds that depend on the population confidence radii of typical trees.

THEOREM 2 (Adaptivity for β-mixing; proven in Appendix B). *Make Assumption 3 in addition to the assumptions of Theorem 1, and consider the typical trees $\mathcal{T}(h, \pi_*)$ with parameters $h \in \mathbb{N}, \pi_* > 0$ such that for $\delta_0 \in (0, 1/e)$*

$$n \ge 2\max\left\{40h, \left\lceil\frac{48\Gamma\mathrm{L}}{\pi_*\delta_0}\right\rceil^{1/\theta}\right\} \times \left\{1+\frac{1200}{\pi_*}\log\Bigl(\frac{24(h+1)}{\delta_0\pi_*}\Bigr)\right\}. \tag{4.2}$$

Then the following inequality holds with probability at least $1-\delta-\delta_0$, simultaneously over all $x \in$ supp:

$$\begin{aligned} &\bigl\|d\bigl(\hat{P}_n(\cdot|x), p(\cdot|x)\bigr)\bigr\|_{\mathrm{L},F} \\ &\quad\le \inf_{T\in\mathcal{T}(h,\pi_*)} \frac{2c+2}{c-1}\bigl\|\gamma(T(x))\bigr\|_{\mathrm{L},F} + (1+2c)\bigl\|\overline{\mathsf{cr}}(T(x))\bigr\|_{\mathrm{L},r}. \end{aligned}$$

Theorem 2 shows that the estimator balances continuity rates and population confidence radii over the set of typical trees. The parameters π_T^{-1} and h_T of these trees may grow polynomially with the sample size n, and this allows for the use of very deep nodes for the estimation of difficult pasts. This strong adaptivity property may be rephrased as an *oracle inequality* when $F = \infty$.

COROLLARY 1 (Oracle inequality). *In the setting of Theorem 2, take any choice (h, π_*) that satisfies (4.2) and set $F = \infty$, $\delta = n^{-a}$ with $a > 0$. Then there*

exists a constant $C > 0$ depending only on the slack parameter $c > 1$, on the alphabet $|A|$ and on the exponent $a > 0$ of δ, such that with probability at least $1 - n^{-a} - \delta_0$,

$$\sup_{\substack{T\in\mathcal{T}(h,\pi_*)\\ \tilde{p}\text{ compatible}\\ \text{with } T}} \left(\sup_{x\in\operatorname{supp}} \frac{\|d(\hat{P}_n(\cdot|x), p(\cdot|x))\|_{\mathrm{L},\infty}}{\|d(\tilde{p}(\cdot|x), p(\cdot|x))\|_{\mathrm{L},\infty} + \|\{\sqrt{\frac{\log n}{\pi_\ell(T(x))n}}\}_{\ell=1}^{\mathrm{L}}\|_{\mathrm{L},\infty}} \right) \le C.$$

This is a consequence of the previous Theorem 2 because any $\tilde{p}$ that is constant on the leaves of T will make errors that are proportional to the continuity rates at those leaves, therefore, adapting its precision to different parts of the tree. Alternatively, we could compute a different estimators for the context tree for each process, namely $\hat{T}_{n,\ell}$ for $\ell = 1, \ldots, \mathrm{L}$. Under the stated conditions, both approaches lead to the same rate of convergence and the pruning rules imply that $\hat{T}_n \subset \bigcup_{\ell=1}^{\mathrm{L}} \hat{T}_{n,\ell}$. The potential advantage of the group approach is to provide a single context tree that is applicable to all processes. However, under different choices of F exploiting the group context tree can lead to improvements as discussed earlier (see examples in Section 5).

The proof of Theorem 2 shows that (4.2) suffices as a requirement for the empirical frequency of any leaf $w \in T$ in the sample to be close to its expected frequency, for any given $T \in \mathcal{T}(h, \pi_*)$. In the next section, we consider important classes of processes that fall within this β-mixing framework.

5. Rates of convergence for theoretical examples. In what follows, we apply the finite sample analysis from the previous section to obtain asymptotic results for some classes of processes. Throughout this section, we assume that $\mathcal{S}$ and A are fixed. The sample size n diverges, and for each n we have parameters $\delta^{(n)}, \delta_0^{(n)}, \mathrm{L}^{(n)}$ and processes

$$X^{(n)}(1), \ldots, X^{(n)}\big(\mathrm{L}^{(n)}\big).$$

We impose the restrictions

$$\delta^{(n)} + \delta_0^{(n)} = O\big(n^{-\xi}\big) \quad \text{and} \quad \mathrm{L}^{(n)}\big(\delta^{(n)}\delta_0^{(n)}\big)^{-1} = O\big(n^{\alpha}\big)$$

for constants $\alpha \ge \xi > 0$. For each example, we make mixing assumptions that we assume to hold uniformly in n and ℓ. We will omit the superscript (n) from our notation.

5.1. *Parametric case.* In our first example, we assume that the true model for the L processes has a finite context tree T^*, which is allowed to vary with n. For a fixed n, this implies the L processes are finite Markov chains, thus exponentially ϕ-mixing; we assume *uniform* exponential β-mixing over all processes and all values of n. We also assume that T^* is a *complete tree*, meaning that any node has 0 or $|A|$ children (cf. Remark 3 in the supplementary material for some comments on this condition which is needed just to achieve uniqueness of the context tree).

EXAMPLE 1 (Parametric case). The processes $X(1), \dots, X(\mathrm{L})$ are stationary and ergodic. Moreover, there exist a finite complete tree T^* and transition probabilities $p = (p_1, \dots, p_{\mathrm{L}})$ that are compatible with the processes

$$\forall 1 \le \ell \le \mathrm{L}, \forall a \in A: \qquad \mathbb{P}\big(X(\ell)_0 = a | X(\ell)_{-\infty}^{-1}\big) = p_\ell\big(a | T^*\big(X(\ell)_{-\infty}^{-1}\big)\big) \qquad \text{a.s.}$$

Moreover, each of these processes is stationary β-mixing with the same exponential rate function:

$$\beta(b) = \chi e^{-\nu b},$$

where $\chi, \nu > 0$ are independent of the sample size. We assume that $h_{T^*}\pi_{T^*}^{-1} = o(n/\log n)$ and $\pi_{T^*}^{-1} = O(n^{1-\varepsilon})$ for some $\varepsilon > 0$.

Finally, we define $d_{T^*} \equiv 1$ if $T^* = \{e\}$; otherwise, when $T^* \neq \{e\}$, we set

$$d_{T^*} \equiv \inf_{w \text{ leaf of } T^*, w \neq e} \Big\{ \sup_{w' \succeq \mathrm{par}(w) \text{ leaf of } T^*} d\big(p(\cdot|w), p\big(\cdot|w'\big)\big) \Big\}.$$

We assume

$$d_{T^*}^{-1} = o\Bigg(\sqrt{\frac{\pi_{T^*} n}{\log n}}\Bigg).$$

Note that, by Remark 3 in the supplementary material [4], $d_{T^*} > 0$ is equivalent to requiring that T^* is the unique minimal complete context tree compatible with the processes $X(1), \dots, X(\mathrm{L})$. Our analysis implies that the "leaf separation quantity" d_{T^*} above is an appropriate detection threshold.

We have the following result.

THEOREM 3. *In the parametric case considered in Example 1, with probability* $1 - O(n^{-\xi})$, *we have* $\hat{T}_n = T^*$ *and*

$$\sup_{x \in \mathrm{supp}} \big\| d\big(\hat{P}_n(\cdot|x), p(\cdot|x)\big) \big\|_{\mathrm{L},F} = O\Bigg(\sqrt{\frac{\log n}{\pi_{T^*} n}}\Bigg).$$

Moreover, the $\log n$ *term in the error estimate may be improved to* $\log\log n$ *in the "many processes" case of Theorem 1.*

Remark 5 in the supplementary material [4] shows that this compares favourably with the theorem of Bühlmann and Wyner [11] for the case $\mathrm{L} = 1$.

5.2. *Chains with infinite connections.* In our second example, we allow for infinite-order chains, but require a nonnullness condition and polynomial uniform continuity.

EXAMPLE 2 (Chains with infinite connections). The processes $X(1), \ldots, X(\mathrm{L})$ are stationary and ergodic. There exist constants $\eta > 0$, $\theta > 1 + 2\alpha$ and $\Gamma_0 > 0$ not depending on the sample size n such that for all $1 \le \ell \le \mathrm{L}$,

$$(\textit{nonnullness}): \qquad \inf_{a \in A, x \in A_{-\infty}^{-1}} p_\ell(a|x) \ge \eta$$

and

$$\forall k \in \mathbb{N} \qquad \max_{w \in A_{-k}^{-1}} \|\gamma(w)\|_{\mathrm{L},F} \le \Gamma_0 k^{-1-\theta}.$$

In this case, we have the following uniform bound.

THEOREM 4. *In the case of chains with infinite connections considered in Example 2, we have*

$$\mathbb{P}\Big(\sup_{x \in \mathrm{supp}} \|d(\hat{P}_n(\cdot|x), p(\cdot|x))\|_{\mathrm{L},F} = O\Big(\frac{1}{\log^{1+\theta} n}\Big)\Big) = 1 - O(n^{-\xi}).$$

This result shows that $\hat{P}_n(\cdot|x)$ converges to $p(\cdot|x)$ uniformly over pasts x, albeit at a slow rate $1/\log^{1+\theta} n$. Section 9 in the supplementary material [4] shows that this is the minimax rate for uniform convergence over pasts, when $\mathrm{L} = 1$ and $A = \{0, 1\}$. Nonetheless, because of the adaptivity of the estimator, faster rates of convergence would be achieved for pasts with better continuity rates.

5.3. *Renewal processes.* Our last example consists of stationary binary renewal processes whose arrival distributions have uniformly bounded $2 + \theta$ moments, $\theta > 0$.

EXAMPLE 3 (Renewal processes). Each process $X(\ell)$ is a stationary and ergodic binary renewal process. The arrival distributions μ_ℓ have support on the whole of $\mathbb{N}$ and satisfy

$$\sum_{k \in \mathbb{N}} \mu_\ell(k) k^{2+\theta} \le C$$

for constants $C, \theta > 0$ that do not depend on $1 \le \ell \le \mathrm{L}$ or on the sample size. Moreover, there exist values $\{f_\ell\}_{\ell=1}^{\mathrm{L}}$ (possibly depending on n) such that

$$f_\ell = \lim_{k \to +\infty} \frac{\mu_\ell(k)}{\sum_{j \ge k} \mu_\ell(j)}. \tag{5.1}$$

In this example, we have no control over the continuity rates of the process at arbitrarily deep levels. We establish the following result.

THEOREM 5. *In the case of renewal processes as in Example 3, let $G \subset A_{-\infty}^{-1}$ be the subset of all strings $x = \cdots 10^{s-1}$ where s is such that*

$$\min_{1\le \ell \le \mathrm{L}} \sum_{j \ge s} \mu_\ell(j) \ge n^{-\frac{\theta}{\theta+1}} \log n.$$

Then the AGCT estimator satisfies the following with probability $1 - O(n^{-\xi})$:

$$\forall x = \cdots 10^{s-1} \in G : \quad \left\| d\big(\hat{P}_n(\cdot|x), p(\cdot|x)\big) \right\|_{\mathrm{L},\infty} \le C \left\| \left\{ \sqrt{\frac{\log n}{n \sum_{j \ge s} \mu_\ell(j)}} \right\}_{\ell=1}^{\mathrm{L}} \right\|_{\mathrm{L},\infty}.$$

Theorem 5 highlights the adaptivity of the rates of convergence. Indeed for pasts $x \in G$ that are more frequent, corresponding to larger values of $\sum_{j \ge s} \mu_\ell(j)$, a faster rate of convergence is obtained.

6. Example of applications to functionals. In this section, we develop two applications of the AGCT model and estimation algorithms. In both cases, the main objects of interest are neither the context trees, nor the transition probabilities, but rather functionals of these quantities. In what follows, we estimate these functionals based on $\hat{T}_n$ and $\hat{P}_n$ accounting for the estimation error and possible misspecification. These two applications rely on different metrics and penalty functions, providing a motivation for the generality of the previous analysis.

6.1. *Discrete stochastic dynamic programming.* Discrete stochastic dynamic programming (DSDP) focuses on solving structured optimization problems in which a control u is chosen from a set of discrete options $\mathcal{U}$ at time t and yields some instantaneous payoff $f(a,u)$, where $a \in A$ is the current state. The system evolves to a state x_{t+1} at period $t+1$ according to an A-valued random function $s(x_{-\infty}^t, u)$ satisfying

$$\mathbb{P}\big(s(x_{-\infty}^t, u) = a\big) = p_u\big(a|x_{-\infty}^t\big) \qquad (a \in A, u \in \mathcal{U}).$$

That is, the transition probabilities of $s(x_{-\infty}^t, u)$ depend on the chosen control $u \in \mathcal{U}$ and (potentially) the complete history of states $x_{-\infty}^t \in A_{-\infty}^{-1}$.

In applications, the main object of interest is the value function that characterize the expected future payoffs as a function of the history of states:

$$V(x) = \max_{u \in \mathcal{U}} \big\{ f(x_{-1}, u) + \lambda \mathbb{E}\big[V\big(xs(x,u)\big)\big] \big\},$$

where $\lambda < 1$ is the discount factor and $xs(x,u)$ is the concatenation of x with $s(x,u)$. In practice, the transition probabilities between states need to be estimated. However, even if transition probabilities were known a priori, the tractability of a dynamic programming formulation relies on avoiding a large state space

(in this case potentially $A_{-\infty}^{-1}$). Nonetheless, the selected state space needs to be rich enough to capture the main features of the transition function $s(\cdot,\cdot)$.

Our motivation to apply the AGCT estimator is to create estimates for the transition probabilities while maintaining a data-driven manageable state space. This is exactly the case in which using the AGCT model can be more attractive than using a (potentially much larger) compatible tree T^*. We advocate in favor of a small approximation error (i.e., comparable with the noise in the estimation) with a substantially smaller state space. Thus, for $x \in A_{-\infty}^{-1}$, we propose to estimate the value function with

$$\hat{V}(x) = \hat{V}\big(\hat{T}_n(x)\big),$$

and the transition probabilities with $\hat{p}_{n,u}(\cdot|\hat{T}_n(x)) = \hat{P}_{n,u}(\cdot|x)$, which are allowed to depend on the action $u \in \mathcal{U}$. The total number of states of the estimated system is the number of leaves of $\hat{T}_n$.

Let the number of groups $\mathrm{L} = |\mathcal{U}|$, $d_\ell = \|\cdot\|_1/2$ and $F = r = \infty$. The data consists of $|\mathcal{U}|$ time series of length n where on each series the decision is chosen to be constant $u \in \mathcal{U}$.

THEOREM 6. *In the discrete stochastic dynamic programming problem, by choosing* $\hat{\mathsf{cr}}$ *as in Definition 1, we have that with probability at least* $1-\delta$ *the estimator* $\hat{V}$ *of the value function satisfies*

$$\sup_{x\in\mathrm{supp}} \frac{|\hat{V}(x)-V(x)|}{\sup_{a\in A}|V(xa)|\inf_{T}\{\|\gamma(T(x))\|_{\mathrm{L},1}+\|\hat{\mathsf{cr}}(T(x))\|_{\mathrm{L},\infty}\}} \le \frac{\lambda}{1-\lambda}4c\frac{c+1}{c-1},$$

where $\|\hat{\mathsf{cr}}(T(x))\|_{\mathrm{L},\infty} \lesssim \max_{\ell=1,\dots,\mathrm{L}} \sqrt{\frac{\log(n\mathrm{L}/\delta)+|A|}{N_{n-1,\ell}(T(x))}}$.

As before, the estimator enjoys adaptivity. In particular, if we restrict the minimum above to typical trees we have the following corollary.

COROLLARY 2 (Value function approximation for β-mixing). *Under the same assumptions of Theorems 2 and 6 with probability at least* $1-\delta-\delta_0$

$$\sup_{T\in\mathcal{T}(h,\pi_*)}\sup_{x\in A_{-\infty}^{-1}} \frac{|\hat{V}(x)-V(x)|}{\sup_{a\in A}|V(xa)|\{\|\gamma(T(x))\|_{\mathrm{L},1}+\max_{\ell=1,\dots,\mathrm{L}}\sqrt{\frac{\log(n/\delta)}{n\pi_\ell(T(x))}}\}} \le \frac{C\lambda}{1-\lambda}.$$

6.2. *Dynamic discrete choice models.* In dynamic discrete choice models, a group of agents makes choices among the same set of options over time [1, 2, 7, 8, 12]. Models usually pre-specify a Markovian structure of the process, which is commonly assumed to be of order 1. We are interested in relaxing this assumption and to estimate the relevant context tree and the associated transition probabilities.

Agents are assumed to be sampled independently from the same population. We assume that the underlying context tree is the same across agents, but allow for the specific transition probability to vary by agent to account for heterogeneity. Herein we focus on the case with no covariates, but results can be extended to the case of discrete covariates [7, 8].

In applications, the main interest is on statistics that are functions of the conditional probabilities rather than the conditional probabilities themselves. Here, we focus on the average marginal dynamic effect for $a \in A$, $x, y \in A_{-\infty}^{-1}$

$$\mathrm{AVEm}(a, x, y) = \mathbb{E}\big[m_\ell(a, x, y)\big],$$

where the marginal dynamic effect $m_\ell(a, x, y) = p_\ell(a|x) - p_\ell(a|y)$, and the expectation is taken over the distribution of agents in the population of interest. The average marginal dynamic effect measures the average over the population of the change in the probability of selection of an option $a \in A$ between two different histories of past consumption $x, y \in A_{-\infty}^{-1}$. Other measures of interest in the literature are the long run proportions of a particular option being chosen, or the probability of selecting a particular option t periods ahead given the current state; see [7].

The estimator of the marginal dynamic effect for an option $a \in A$ and histories of consumptions $x, y \in A_{-\infty}^{-1}$ for the ℓth agent is

$$\hat{m}_\ell(a, x, y) = \hat{p}_{n,\ell}\big(a|\hat{T}_n(x)\big) - \hat{p}_{n,\ell}\big(a|\hat{T}_n(y)\big),$$

and the estimator for the average marginal dynamic effect is

$$\widehat{\mathrm{AVEm}}(a, x, y) = \frac{1}{\mathrm{L}} \sum_{\ell=1}^{\mathrm{L}} \hat{m}_\ell(a, x, y).$$

Therefore, if the conditional probabilities were known, a rate of $1/\sqrt{L}$ would be optimal for the estimation of a single average marginal dynamic effect. In what follows, we will use the AGCT model to estimate these dynamic effects uniformly over all histories. This motivates the choice of $d_\ell = \|\cdot\|_\infty$, $F = 1$ and $r = 2$ in the AGCT estimator.

THEOREM 7. *In the dynamic discrete choice model, if the context tree and conditional probabilities are estimated with* $\hat{\mathsf{cr}}$ *as in Definition 2, we have that with probability at least* $1 - 2\delta$ *the estimator for the average marginal dynamic effect satisfies*

$$\sup_{\substack{a\in A,\\ x,y\in\mathrm{supp}}} \frac{|\widehat{\mathrm{AVEm}}(a, x, y) - \mathrm{AVEm}(a, x, y)|}{\max_{z=x,y} \inf_T \{\|\gamma(T(z))\|_{\mathrm{L},1} + \|\hat{\mathsf{cr}}(T(z))\|_{\mathrm{L},2}\} + \sqrt{\frac{2\log(\frac{|A|\cdot n^4}{4\delta})}{\mathrm{L}}} + \frac{2}{\mathrm{L}}} \le 8c\frac{c+1}{c-1},$$

where $\|\hat{\mathsf{cr}}(T(z))\|_{\mathrm{L},2} \lesssim \sqrt{\frac{\log\log n + \log|A|}{\mathrm{L}} \sum_{\ell=1}^{\mathrm{L}} 1/N_{n-1,\ell}(T(z))}$, $z \in A_{-\infty}^{-1}$.

This uniform rate of convergence for the average marginal dynamic effect is governed by the rate of convergence of the conditional probabilities of the best context tree estimator, and the number of different agents in the data. Interestingly, the above result holds uniformly over all pairs $x, y \in A_{-\infty}^{-1}$.

7. Linguistic rhythm differences between European and Brazilian Portuguese. In this section, we revisit the application and the data considered in [17] regarding the linguistic features underlying the European Portuguese (EP) and Brazilian Portuguese (BP) languages. The goal of [17] was to compare the rhythmic fingerprints of the two languages in written form.

For each language, the data consist of articles from a popular daily newspaper from the years 1994 and 1995. For each year and each newspaper, 20 articles were randomly selected. The linguistic features are represented by a quinary alphabet with four rhythmic features (0, 1, 2, 3) and an additional feature representing the end of an article (4). The four rhythmic features represent: nonstressed, nonprosodic word initial syllable (0); stressed, nonprosodic word initial syllable (1); no-stressed, prosodic word initial syllable (2); and stressed prosodic word initial syllable (3). Each data sample was then treated as a stochastic process, and a variant of the BIC model selection method was used to fit a context tree to each sample. Their main finding was summarized as follows.

> [T]he main difference between the two languages is that whereas in BP both 2 (unstressed boundary of a phonological word) and 3 (stressed boundary of a phonological word) are contexts, in EP only 3 is a context. This means that in EP, as far as noninitial stress words are concerned, the choice of lexical items is dependent on the rhythmic properties of the preceding words. This is not true when the word begins with a stressed syllable. This does not occur in BP, where word boundaries are always contexts, and as such insensitive to what occurs before, independently of being stressed or not. These statistical findings are compatible with the current discussion in the linguistic literature concerning the different behavior of phonological words in the two languages [...] (Galves et al., [17], Section 6)

In [17], for each newspaper, the 40 days sample is concatenated into a single string containing respectively a sequence of 105,326 and 97,750 linguistic features. In order to concatenate articles from different days, a homogeneity assumption was required. However, heterogeneity over different days, or at least over the different years are a source of potential concern. For example, 1994 was a World Cup year and the media in both countries are heavily influenced by such event. Our own study accounts for possible heterogeneity on the conditional probabilities by treating each year as a group in the group context tree model. Thus, we allow for year specific conditional probabilities.

Figure 1 displays the estimated context trees. Our findings are in good agreement with [17], in that the context trees found for BP in both studies are the same, and our tree for EP strictly contains the one found in [17]. In particular, we corroborate their finding that 2 is a context for BP but *not* for EP.

8. Discussions and variations.

8.1. *Comparisons with single-process case.* We briefly indicate similarities and differences between the results presented above with [16], which concerns the single-process case. The work [16] proves weak consistency in the estimation of

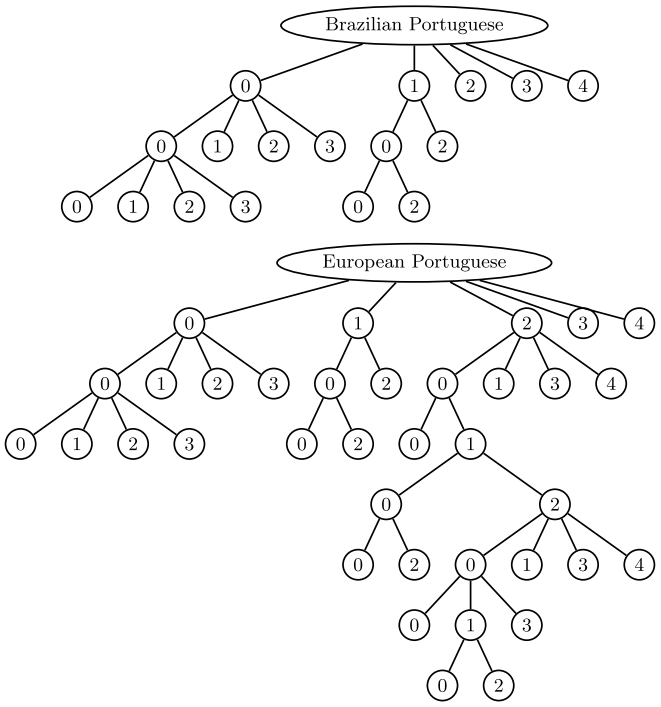

FIG. 1. *Estimated context trees for the Brazilian Portuguese and European Portuguese languages based accounting for heterogeneity in different years.*

conditional probabilities and of (truncated) context trees for all nodes in a tree T_n that grows with the sample size n. For this, they assume that the stochastic process is geometrically α-mixing, and also that there is sufficient separation between the conditional probabilities corresponding to leaves of the tree and their parents. The authors of [16] point out that the latter assumptions might be hard to check in practice.

Our analysis differs from theirs in several important aspects even in the case of $\mathrm{L} = 1$ processes. Our goal is to estimate transition probabilities given the entire infinite past, uniformly over all such pasts. Achieving consistency in our setting requires that these probabilities be continuous functions of the infinite past, which [16] do not need to assume. By contrast, given continuity and β-mixing, model selection and probability estimation become separate tasks. In particular, our results on the transition probabilities do not require any kind of separation between leaves and their parents. In addition, our results cover natural and interesting classes of processes (such as certain renewal processes) where geometric mixing bounds are not available. Other points of the analysis are mostly incomparable due to the differences in assumptions.

8.2. *Computational efficiency and variations.* The algorithm can be implemented efficiently, that is, in polynomial time with respect to the parameters L and n of the data. Observe that $\mathsf{CanRmv}(w)$ can be computed efficiently from the list of values:

$$\mathrm{List}(w) \equiv \big\{\big(\hat{p}_n(\cdot|w'), \hat{\mathsf{cr}}(w')\big) : w' \in E_n, w' \succeq w\big\}$$

and the corresponding list for $\mathrm{par}(w)$. Since $\mathsf{CanRmv}(w)$ is only computed for leaves of the current tree $\hat{T}_n$, we only need to ensure that at all times, each leaf node and each parent of a leaf stores the correct list $\mathrm{List}(w)$. This can be achieved as follows:

- initially, one sets $\mathrm{List}(w) = \{\hat{p}_n(\cdot|w), \hat{\mathsf{cr}}(w)\}$ for each $w \in E_n$;
- whenever a leaf w is examined in $\hat{T}_n$, its parent's list is updated:

$$\mathrm{List}\big(\mathrm{par}(w)\big) \leftarrow \mathrm{List}\big(\mathrm{par}(w)\big) \cup \Bigg(\bigcup_{w' \in \hat{T}_n : \mathrm{par}(w') = \mathrm{par}(w)} \mathrm{List}\big(w'\big) \Bigg).$$

  Actually, this update only needs to be performed at the first time a child of $\mathrm{par}(w)$ is examined.

We note in passing that more efficient algorithms can be found for the case $\mathrm{L} = 1$ with the ℓ_∞ metric by using compact suffix trees. This will be elaborated upon in a companion paper.

All results established in this work would remain valid if in the definition of $\mathsf{CanRmv}(w)$ in (2.4) we set $w'' \in \mathcal{W}$ where $\mathrm{par}(w) \in \mathcal{W} \subseteq \{z \in E_n : z \succeq \mathrm{par}(w)\}$. For the same choice of confidence radius, computationally we would like to use the smallest set $\mathcal{W}$ while statistically we would like to use the largest such set.

8.3. *Improvement on confidence radii based on maximal variance.* The choices of confidence radii described in Definitions 1 and 2 do not explore the intrinsic variance within the norm, namely

$$\bar{\sigma}_\ell^2(w) := \max_{S \in \mathcal{S}} \bar{p}_{n,\ell}(S|w)\big(1 - \bar{p}_{n,\ell}(S|w)\big)$$

where $\bar{p}_{n,\ell}(S|w)$ is a weighted sum of probabilities defined in (A.3) for which $\hat{p}_{n,\ell}(S|w)$ is a consistent estimator. (These probabilities can be seen as an oracle estimator; see Section 1 in the supplementary material [4] for a discussion.) Generically, adding variance to our bounds does not necessarily improve rates of convergence but can improve finite sample performance, particularly in the case of $d_\ell = \|\cdot\|_\infty$ with $|A| > 2$. Here, we discuss such a modification of Definition 1 that leads to strictly smaller confidence radii while still achieving the same guarantees as in Theorem 1. However, the variance-based control can be applied to a suffix w only if there were enough occurrences of the suffix in the data, namely the following event occurred:

$$J_{\ell,w} := \left\{ N_{n-1,\ell}(w) \geq \frac{2\log(n^2|\mathcal{S}|/\delta) + 4\log[2 + 2\log(\bar{\sigma}_\ell^2(w) N_{n-1,\ell}(w))]}{\bar{\sigma}_\ell^2(w)\log^2(3/2)} \right\}.$$

Otherwise, we use the previous choice as in Definition 1. To concisely state the results regarding the maximum variance, we define

$$\tilde{\sigma}_\ell(w) := \sqrt{2}\bar{\sigma}_\ell(w)1_{\{J_{\ell,w}\}} + 1_{\{J_{\ell,w}^c\}} \le 1.$$

We define $\hat{\mathsf{cr}}_\ell^{\tilde{\sigma}}(w) := \tilde{\sigma}_\ell(w)\hat{\mathsf{cr}}_\ell(w)$. By construction, it follows that $\hat{\mathsf{cr}}_\ell^{\tilde{\sigma}}(w) \le \hat{\mathsf{cr}}_{\ell,m}(w)$ since $\tilde{\sigma}_\ell(w) \le 1$. However, $\hat{\mathsf{cr}}_\ell^{\tilde{\sigma}}(w)$ might not be nonincreasing in w. Nonetheless, the confidence radius $\hat{\mathsf{cr}}_\ell^{\tilde{\sigma}}(w)$ can be majorated by the monotone confidence radius which still leads to an improvement over $\hat{\mathsf{cr}}_{\ell,m}(w)$, namely

$$\hat{\mathsf{cr}}_\ell^*(w) = \max_{w' \preceq w} \hat{\mathsf{cr}}_\ell^{\tilde{\sigma}}(w') \le \max_{w' \preceq w} \hat{\mathsf{cr}}_{\ell,m}(w') = \hat{\mathsf{cr}}_{\ell,m}(w).$$

A side remark is that $\hat{\mathsf{cr}}_\ell^*(w)$ requires the estimation of $\bar{\sigma}_\ell(w)$. Indeed, the estimates need to satisfy $\bar{\sigma}_\ell(w) \le \hat{\sigma}_\ell(w)$ with high probability uniformly over $w \in E_n$. However, it follows that any such estimator will satisfy $\hat{\sigma}_\ell(w) \le 1/2$ so that even by setting $\hat{\sigma}_\ell(w) = 1/2$ we still achieve smaller confidence radius than the original definition.

9. Conclusion. Understanding the memory structure of stochastic processes has proved to be of fundamental importance in applications. VLMC models have been playing a central role in modeling and estimating stationary processes with discrete alphabets. In this work, we consider an extension of the traditional VLMC in which many stationary processes share the same context tree but potentially different conditional probabilities. Since we allow for potentially infinite memory processes, we propose to focus the estimation on an oracle context tree that optimally balances the bias and variance trade-off for a given sample.

We propose a computationally efficient estimator for the underlying context tree and the associated conditional probabilities. We establish several properties of the proposed estimator, including adaptivity and oracle inequalities for the estimation of conditional probabilities. We propose and analyze data-driven choices of the penalty parameters for the regularization, and study its typical behavior under β-mixing conditions. Two applications, discrete dynamic stochastic programming and discrete choice models, motivated the proposal of the AGCT model. In these applications, we are interest in functionals of the conditional probabilities. We developed the uniform bounds for the estimation of these functionals accounting for possible misspecification of the estimated context tree.

Finally, we investigate the application of the group context tree model and the proposed estimators to investigate the rhythmic differences between Brazilian and European Portuguese allowing for possible heterogeneity in the sample. Our results fully support previous findings of the literature.

APPENDIX A: PROOF OF THEOREM 1

Theorem 1 follows directly from three lemmas related to the *good event* Good_*:

$$\mathrm{Good}_* \equiv \bigcap_{w\in A^*} \left\{ \left\| d\big(p(\cdot|w), \hat{p}_n(\cdot|w)\big)\right\|_{\mathrm{L},F} \le \left\| \{\gamma_\ell(w)\}_{\ell=1}^{\mathrm{L}} \right\|_{\mathrm{L},F} + \left\| \hat{\mathsf{cr}}(w)\right\|_{\mathrm{L},r} \right\}, \tag{A.1}$$

where $r = F \in [1, +\infty]$ in the "general case" and $r = 2$, $F = 1$ in the "many processes" case of Theorem 1, and for $|w| < \infty$ we define $p_\ell(a|w) = \mathbb{P}(X_0(\ell) = a | X_{-|w|}^{-1}(\ell) = w)$, if $\mathbb{P}(X_{-|w|}^{-1}(\ell) = w) > 0$, and $p_\ell(a|w) = 1/|A|$ if $\mathbb{P}(X_{-|w|}^{-1}(\ell) = w) = 0$.

LEMMA 1 (Proven in Section A.1). *If* Good_* *holds*, $\hat{T}_n \subset T^*$.

LEMMA 2 (Proven in Section A.2). *If* Good_* *holds, then for all* $x \in A_{-\infty}^{-1}$ *and any finite tree* T

$$\left\| d\big(p(\cdot|x), \hat{P}_n(\cdot|x)\big)\right\|_{\mathrm{L},F} \le \frac{2c+2}{c-1} \left\| \{\gamma_\ell(T(x))\}_{\ell=1}^{\mathrm{L}} \right\|_{\mathrm{L},F} + (1+2c) \left\| \hat{\mathsf{cr}}(T(x))\right\|_{\mathrm{L},r}.$$

LEMMA 3 (Proven in Section A.3). *The probability of* Good_* *is* $\ge 1-\delta$.

These three lemmas are proven subsequently.

A.1. Proof of Lemma 1. Let $z \in A^* \setminus T^*$ and assume Good_*; we will show that $z \notin \hat{T}_n$. Let w be an ancestor of z which is a leaf of T^*. Because T^* is the true context tree, $\|\{\gamma_\ell(w')\}_{\ell=1}^{\mathrm{L}}\|_{\mathrm{L},F} = 0$ for all descendants of w, in particular for z, $\mathrm{par}(z)$ and their descendants. If we assume Good_* holds, the triangle inequality gives

$$\forall u, v \succeq \mathrm{par}(z) : \left\| d\big(\hat{p}_n(\cdot|u), \hat{p}_n(\cdot|v)\big)\right\|_{\mathrm{L},F} \le \left\| \hat{\mathsf{cr}}(u)\right\|_{\mathrm{L},r} + \left\| \hat{\mathsf{cr}}(v)\right\|_{\mathrm{L},r},$$

and one can easily deduce from this that $\mathrm{CanRmv}(u) = 1$ for all $u \succeq z$. This means z is pruned from the tree.

A.2. Proof of Lemma 2. Fix x and T. Recall that $\hat{P}_n(\cdot|x) = \hat{p}_n(\cdot|\hat{T}_n(x))$ and that $\|\hat{\mathsf{cr}}(w)\|_{\mathrm{L},r}$ is monotone nondecreasing in w. Notice that $\hat{T}_n(x)$ and $T(x)$ are both finite suffixes of x. This allows us to divide the analysis into three cases.

Case 0: $\hat{T}_n(x) = T(x)$. The result follows from

$$\begin{aligned}
&\left\| d\big(p(\cdot|x), \hat{p}_n\big(\cdot|\hat{T}_n(x)\big)\big)\right\|_{\mathrm{L},F} \\
&\quad = \left\| d\big(p(\cdot|x), \hat{p}_n\big(\cdot|T(x)\big)\big)\right\|_{\mathrm{L},F} \\
&\quad \le \left\| d\big(p(\cdot|x), p\big(\cdot|T(x)\big)\big)\right\|_{\mathrm{L},F} + \left\| d\big(p\big(\cdot|T(x)\big), \hat{p}_n\big(\cdot|\hat{T}_n(x)\big)\big)\right\|_{\mathrm{L},F} \\
&\quad \le 2\left\| \{\gamma_\ell(T(x))\}_{\ell=1}^{\mathrm{L}} \right\|_{\mathrm{L},F} + \left\| \hat{\mathsf{cr}}(T(x))\right\|_{\mathrm{L},r},
\end{aligned}$$

where the first equality is from $\hat{T}_n(x) = T(x)$, the second step from triangle inequality, and the third from the event Good_* and the definition of the continuity rates.

Case 1: $\hat{T}_n(x) \prec T(x)$. Let w denote the child of $\hat{T}_n(x)$ on the path to $T(x)$. Note that w must have been pruned, otherwise $w \in \hat{T}_n$ would be a longer suffix of x than $\hat{T}_n(x)$.

We deduce that w satisfies $\mathsf{CanRmv}(w) = 1$, like any other pruned node. In particular, this implies that $T(x) \succeq w$ and $\hat{T}_n(x) = \text{par}(w)$ satisfy

$$\big\|d\big(\hat{p}_n(\cdot|T(x)),\, \hat{p}_n(\cdot|\hat{T}_n(x))\big)\big\|_{\mathrm{L},F} \le c\big[\big\|\hat{\mathsf{cr}}\big(T(x)\big)\big\|_{\mathrm{L},r} + \big\|\hat{\mathsf{cr}}\big(\hat{T}_n(x)\big)\big\|_{\mathrm{L},r}\big].$$

Since $\hat{T}_n(x) \prec T(x)$, the RHS of the above display is $\le 2c\|\hat{\mathsf{cr}}(T(x))\|_{\mathrm{L},r}$, and the occurrence of Good_* gives

$$\big\|d\big(p(\cdot|T(x)),\, \hat{p}_n(\cdot|T(x))\big)\big\|_{\mathrm{L},F} \le \big\|\{\gamma_\ell(T(x))\}_{\ell=1}^{\mathrm{L}}\big\|_{\mathrm{L},F} + \big\|\hat{\mathsf{cr}}\big(T(x)\big)\big\|_{\mathrm{L},r}.$$

Combining these observations and employing the triangle inequality gives

$$\big\|d\big(p(\cdot|T(x)),\, \hat{p}_n(\cdot|\hat{T}_n(x))\big)\big\|_{\mathrm{L},F} \le \big\|\{\gamma_\ell(T(x))\}_{\ell=1}^{\mathrm{L}}\big\|_{\mathrm{L},F} + (1+2c)\big\|\hat{\mathsf{cr}}\big(T(x)\big)\big\|_{\mathrm{L},r}.$$

Using $\|d(p(\cdot|x),\, p(\cdot|T(x)))\|_{\mathrm{L},F} \le \|\{\gamma_\ell(T(x))\}_{\ell=1}^{\mathrm{L}}\|_{\mathrm{L},F}$ and another application of the triangle inequality completes the proof in this case.

Case 2: $\hat{T}_n(x) \succ T(x)$. We make the following claim.

CLAIM 1 (Proven subsequently).

$$\big\|\hat{\mathsf{cr}}\big(\hat{T}_n(x)\big)\big\|_{\mathrm{L},r} + \big\|\hat{\mathsf{cr}}\big(T(x)\big)\big\|_{\mathrm{L},r} \le \frac{3}{c-1}\big\|\{\gamma_\ell(T(x))\}_{\ell=1}^{\mathrm{L}}\big\|_{\mathrm{L},F}.$$

To see how the claim implies the result, we note that

$$\begin{aligned}
\big\|d\big(p(\cdot|x),\, \hat{p}_n(\cdot|\hat{T}_n(x))\big)\big\|_{\mathrm{L},F} &\le \big\|d\big(p(\cdot|x),\, p(\cdot|\hat{T}_n(x))\big)\big\|_{\mathrm{L},F} \\
&\quad + \big\|d\big(p(\cdot|\hat{T}_n(x)),\, \hat{p}_n(\cdot|\hat{T}_n(x))\big)\big\|_{\mathrm{L},F}, \\
\text{(use continuity rates)} &\le \big\|\{\gamma_\ell(\hat{T}_n(x))\}_{\ell=1}^{\mathrm{L}}\big\|_{\mathrm{L},F} \\
&\quad + \big\|d\big(p(\cdot|\hat{T}_n(x)),\, \hat{p}_n(\cdot|\hat{T}_n(x))\big)\big\|_{\mathrm{L},F}, \\
\text{(Good_* holds)} &\le 2\big\|\{\gamma_\ell(\hat{T}_n(x))\}_{\ell=1}^{\mathrm{L}}\big\|_{\mathrm{L},F} \\
&\quad + \big\|\hat{\mathsf{cr}}\big(\hat{T}_n(x)\big)\big\|_{\mathrm{L},r}, \\
\big(T(x) \preceq \hat{T}_n(x) \Rightarrow \gamma_\ell(T(x)) \text{ larger}\big) &\le 2\big\|\{\gamma_\ell(T(x))\}_{\ell=1}^{\mathrm{L}}\big\|_{\mathrm{L},F} \\
&\quad + \big\|\hat{\mathsf{cr}}\big(\hat{T}_n(x)\big)\big\|_{\mathrm{L},r}, \\
\text{(use Claim)} &\le \left(2 + \frac{3}{c-1}\right)\big\|\{\gamma_\ell(T(x))\}_{\ell=1}^{\mathrm{L}}\big\|_{\mathrm{L},F}.
\end{aligned}$$

It remains to prove the claim. Since $\hat{T}_n(x)$ was not pruned, there exist $w' \succeq \hat{T}_n(x)$, $w'' \succeq \mathrm{par}(\hat{T}_n(x)) \succeq T(x)$ with

$$c[\|\hat{\mathsf{cr}}(w')\|_{\mathrm{L},r} + \|\hat{\mathsf{cr}}(w'')\|_{\mathrm{L},r}] < \|d(\hat{p}_n(\cdot|w'), \hat{p}_n(\cdot|w''))\|_{\mathrm{L},F}. \tag{A.2}$$

On the other hand,

$$\begin{aligned}\|d(\hat{p}(\cdot|w'), \hat{p}_n(\cdot|w''))\|_{\mathrm{L},F} \le{}& \|d(p(\cdot|w'), p(\cdot|w''))\|_{\mathrm{L},F} \\ &+ \|d(\hat{p}(\cdot|w'), p(\cdot|w'))\|_{\mathrm{L},F} \\ &+ \|d(\hat{p}(\cdot|w''), p(\cdot|w''))\|_{\mathrm{L},F}.\end{aligned}$$

The first term in the RHS is $\le \|\{\gamma_\ell(T(x))\}_{\ell=1}^{\mathrm{L}}\|_{\mathrm{L},F}$ since $w', w'' \succeq T(x)$. The other two terms can be bounded via Good_*, and we obtain

$$\begin{aligned}\|d(\hat{p}(\cdot|w'), \hat{p}_n(\cdot|w''))\|_{\mathrm{L},F} \le{}& 3\|\{\gamma_\ell(T(x))\}_{\ell=1}^{\mathrm{L}}\|_{\mathrm{L},F} \\ &+ \|\hat{\mathsf{cr}}(w')\|_{\mathrm{L},r} + \|\hat{\mathsf{cr}}(w'')\|_{\mathrm{L},r}.\end{aligned}$$

Combining this with (A.2) gives

$$\|\hat{\mathsf{cr}}(w')\|_{\mathrm{L},r} + \|\hat{\mathsf{cr}}(w'')\|_{\mathrm{L},r} \le \frac{3}{c-1}\|\{\gamma_\ell(T(x))\}_{\ell=1}^{\mathrm{L}}\|_{\mathrm{L},F}.$$

The proof of the claim is complete once we recall that $w' \succeq \hat{T}_n(x)$, $w'' \succeq T(x)$ and the confidence radii $\|\hat{\mathsf{cr}}(w)\|_{\mathrm{L},r}$ are monotone functions of w.

A.3. Proof of Lemma 3. We define what one might call *oracle transition probabilities*: given a context $w \in E_n$, $a \in A$, $\ell = 1, \dots, \mathrm{L}$ as

$$\overline{p}_{n,\ell}(a|w) \equiv \frac{1}{N_{n-1,\ell}(w)} \sum_{i=|w|+1}^{n} 1_{\{X_{i-|w|}^{i-1}(\ell)=w\}} p_\ell\big(a|X_{-\infty}^{i-1}(\ell)\big), \tag{A.3}$$

and as $\overline{p}_{n,\ell}(a|w) \equiv 1/|A|$ if $w \notin E_n$. A salient feature is that these random transition probabilities are always close to the actual transition probabilities in the following sense:

$$\text{If } T(x) \in E_n, \qquad \|d\big(p(\cdot|x), \overline{p}_n(\cdot|T(x))\big)\|_{\mathrm{L},F} \le \|\{\gamma_\ell\big(T(x)\big)\}_{\ell=1}^{\mathrm{L}}\|_{\mathrm{L},F}. \tag{A.4}$$

This follows from the fact that $\overline{p}_{n,\ell}(\cdot|T(x))$ is a convex combination of transition probabilities $p_\ell(\cdot|y)$ with $y \succeq T(x)$.

To continue, we choose a parameter $m = \infty$ in the "general case" of Theorem 1, and $m = 2$ in the "many processes" case of the same theorem. The following regularization event will be important in our analysis:

$$\mathrm{Good}_m \equiv \bigcap_{w \in E_n} \left\{ \left\| \left\{ \frac{d_\ell(\overline{p}_{n,\ell}(\cdot|w), \hat{p}_{n,\ell}(\cdot|w))}{\hat{\mathsf{cr}}_\ell(w)} \right\}_{\ell=1}^{\mathrm{L}} \right\|_{\mathrm{L},m} \le 1 \right\}. \tag{A.5}$$

CLAIM 2. $\text{Good}_m \subset \text{Good}_*$, *where* Good_* *was defined in* (A.1).

PROOF. By (A.4) and the triangle inequality, it suffices to show that we have the inequality

$$\big\|d\big(\hat{p}_n(\cdot|w), \bar{p}_n(\cdot|w)\big)\big\|_{\mathrm{L},F} \le \big\|\hat{\mathsf{cr}}(w)\big\|_{\mathrm{L},r}$$

for all $w \in A^*$ whenever Good_m holds. This is trivially true when $w \notin E_n$. When $w \in E_n$, Hölder's inequality implies

$$\big\|d\big(\hat{p}_n(\cdot|w), \bar{p}_n(\cdot|w)\big)\big\|_{\mathrm{L},F} \le \left\|\left\{\frac{d_\ell(\overline{p}_{n,\ell}(\cdot|w), \hat{p}_{n,\ell}(\cdot|w))}{\hat{\mathsf{cr}}_\ell(w)}\right\}_{\ell=1}^{\mathrm{L}}\right\|_{\mathrm{L},m} \big\|\hat{\mathsf{cr}}(w)\big\|_{\mathrm{L},r}$$

and the first term in the RHS is ≤ 1 in Good_m. □

The remainder of the proof consists of showing the following.

CLAIM 3. $\mathbb{P}(\text{Good}_m) \ge 1 - \delta$.

This clearly suffices to complete the proof in both cases.

We will use a martingale framework from Section 7 in the supplementary material [4]. The following is the special case $\gamma = 2$ and $i_0 = \log_2 n$ of Lemma 4 in the Supplementary Material.

LEMMA 4. *Let* $(M_j, \mathcal{F}_j)_{m=0}^n$ *be a martingale with* $M_0 = 0$. *Assume that for each* $1 \le j \le n$ *we have a* $\mathcal{F}_{j-1}$*-measurable indicator random variable* Y_{j-1} *with* $|M_j - M_{j-1}| \le Y_{j-1}$ *almost surely, and define* $V_n \equiv \sum_{j=0}^{n-1} Y_j^2$. *Then*

$$\forall t \ge 0: \qquad \mathbb{P}\left(\frac{M_n^2}{4V_n} - 2\ln(2 + \log_2 V_n) \ge t \,|\, V_n > 0\right) \le e^{-t}.$$

Recall that the metric $d = d_1 = \cdots = d_{\mathrm{L}}$ is given by

$$d_\ell(p, q) = d_{\mathcal{S}}(p, q) = \sup_{A \in \mathcal{S}} |p(A) - q(A)|.$$

We will consider a family of martingales indexed by $w \in A^*$, $S \in \mathcal{S}$ and $1 \le \ell \le \mathrm{L}$. A simple calculation reveals that for any $j \in \mathbb{N}$

$$M_{j,\ell}(wS) = N_{j-1,\ell}(w) \sum_{a \in S} \big(\hat{p}_{j,\ell}(a|w) - \overline{p}_{j,\ell}(a|w)\big)$$

is a martingale under the natural filtration. One may take

$$Y_{j-1} = 0 \qquad \text{if } j - 1 < |w| \quad \text{and} \quad Y_{j-1} = 1\{X_{m-|w|}^{m-1} = w\} \qquad \text{otherwise,}$$

so that the corresponding $V_n = V_{n,\ell}(wS)$ equals $N_{n-1,\ell}(w) \vee 1$. We also have that

$$N_{n-1,\ell}(w) d_\ell\big(\hat{p}_{n,\ell}(\cdot|w), \overline{p}_{n,\ell}(\cdot|w)\big)^2 = \max_{S \in \mathcal{S}} \frac{M_n(wS)^2}{V_n(wS)}$$

whenever $N_{n-1,\ell}(w) > 0$. The following is immediate from this discussion combined with Lemma 4.

LEMMA 5. *For any $w \in A^*$ and any process $1 \le \ell \le \mathrm{L}$, we have*

$\forall t \ge 0$:

$$\mathbb{P}\Big(\frac{N_{n-1,\ell}(w) d_\ell(\hat{p}_{n,\ell}(\cdot|w), \overline{p}_{n,\ell}(\cdot|w))^2}{4[2\ln(2+\log_2 N_{n-1,\ell}(w)) + \ln|\mathcal{S}| + t]} > 1 \Big| N_{n-1,\ell}(w) > 0\Big) \le e^{-t}.$$

We use this lemma to prove Claim 3 in the two cases.

PROOF OF CLAIM 3 IN THE "GENERAL CASE". Set $t := \ln(n^2\mathrm{L}/\delta)$. For $\ell = 1, 2, \ldots, \mathrm{L}$, define

$$A_\ell := \Big\{\frac{N_{n-1,\ell}(w) d_\ell(\hat{p}_n(\cdot|w), \overline{p}_n(\cdot|w))^2}{4[2\ln(2+\log_2 N_{n-1,\ell}(w)) + \ln|\mathcal{S}| + t]} > 1\Big\};$$
$$B_\ell := \{N_{n-1,\ell}(w) > 0\}.$$

Lemma 5 gives $\mathbb{P}(A_\ell|B_\ell) \le \delta/n^2\mathrm{L}$ for each $1 \le \ell \le \mathrm{L}$. Recalling the formula for $\hat{\mathsf{cr}}_\ell(w)$ in Definition 1, we see that $d_\ell(\hat{p}_{n,\ell}(\cdot|w), \overline{p}_{n,\ell}(\cdot|w)) > \hat{\mathsf{cr}}_\ell(w)$ if and only if A_ℓ holds. Therefore,

$$\begin{aligned}
&\mathbb{P}\Big(\Big\|\Big\{\frac{d_\ell(\hat{p}_n(\cdot|w), \overline{p}_n(\cdot|w))}{\hat{\mathsf{cr}}_\ell(w)}\Big\}_{\ell=1}^{\mathrm{L}}\Big\|_{\mathrm{L},\infty} > 1 \Big| \min_\ell N_{n-1,\ell}(w) > 0\Big) \\
&\quad = \mathbb{P}\Big(\bigcup_{\ell=1}^{\mathrm{L}} A_\ell \Big| \bigcap_{\ell'=1}^{\mathrm{L}} B_{\ell'}\Big) \qquad \text{(A.6)} \\
&\quad \le \sum_{\ell=1}^{\mathrm{L}} \mathbb{P}\Big(A_\ell \Big| \bigcap_{\ell'=1}^{\mathrm{L}} B_{\ell'}\Big).
\end{aligned}$$

Now recall that the L processes $X(\ell)$ are all independent, therefore, A_ℓ depends on B_ℓ but not on $B_{\ell'}$ for $\ell' \ne \ell$. We obtain

$$\mathbb{P}\Big(A_\ell \Big| \bigcap_{\ell=1}^{\mathrm{L}} B_{\ell'}\Big) = \mathbb{P}(A_\ell|B_\ell) \le \frac{\delta}{n^2\mathrm{L}}.$$

Plugging this back into (A.6) and removing the conditioning gives

$$\mathbb{P}\Big(\Big\|\Big\{\frac{d_\ell(\hat{p}_n(\cdot|w), \overline{p}_n(\cdot|w))}{\hat{\mathsf{cr}}_\ell(w)}\Big\}_{\ell=1}^{\mathrm{L}}\Big\|_{\mathrm{L},\infty} > 1\Big) \le \frac{\delta}{n^2}\mathbb{P}\Big(\min_{1\le\ell\le\mathrm{L}} N_{n-1,\ell}(w) > 0\Big).$$

Taking a union bound over all $w \in A^*$ and bounding

$$\mathbb{P}\Big(\min_{1\le\ell\le\mathrm{L}} N_{n-1,\ell}(w) > 0\Big) \le \mathbb{P}\big(N_{n-1,1}(w) > 0\big)$$

gives

$$1 - \mathbb{P}(\text{Good}_\infty) \le \frac{\delta}{n^2} \sum_{w \in A^*} \mathbb{P}\big(N_{n-1,\ell}(w) > 0\big).$$

The sum of probabilities in the RHS is the expected number of distinct substrings of $X_1^{n-1}(1)$, which is at most n^2. This implies Good_∞ occurs with probability at least $1 - \delta$, as desired. □

PROOF OF CLAIM 3 IN THE "MANY PROCESSES" CASE. For each $w \in A^*$ and $1 \le \ell \le \mathrm{L}$, define the random variable

$$\begin{aligned}\Delta_\ell(w) \equiv{}& N_{n-1,\ell}(w) d_\ell\big(\hat{p}_{n,\ell}(\cdot|w), \overline{p}_{n,\ell}(\cdot|w)\big)^2 \\ &- \big[4\ln\big(2 + 2\log_2 N_{n-1,\ell}(w)\big) + \ln|\mathcal{S}|\big].\end{aligned} \tag{A.7}$$

The definition of $\hat{\mathsf{cr}}_\ell(w)$ in Definition 2 implies

$$\left\| \left\{ \frac{d_\ell(\hat{p}_{n,\ell}(\cdot|w), \overline{p}_{n,\ell}(\cdot|w))}{\hat{\mathsf{cr}}_\ell(w)} \right\}_{\ell=1}^{\mathrm{L}} \right\|_{\mathrm{L},2} > 1 \quad \Leftrightarrow \quad \frac{1}{\mathrm{L}} \sum_{\ell=1}^{\mathrm{L}} \Delta_\ell(w) > 1 + \sqrt{\frac{6\ln(n^2/\delta)}{\mathrm{L}}}.$$

Lemma 5 implies that, conditionally on $N_{n-1,\ell}(w) > 0$, $\Delta_\ell(w)$ is dominated by an exponential random variable with mean 1. The independence of the L processes implies that

$$\mathbb{P}\left(\frac{1}{\mathrm{L}} \sum_{\ell=1}^{L} \Delta_\ell(w) > 1 + \sqrt{\frac{6\ln(n^2/\delta)}{\mathrm{L}}} \,\Big|\, \min_\ell N_{n-1,\ell}(w) > 0 \right)$$

can be upper bounded as if the $\Delta_\ell(w)$'s were independent exponentials. A standard Laplace transform calculation implies

$$\mathbb{P}\left(\frac{1}{\mathrm{L}} \sum_{\ell=1}^{L} \Delta_\ell(w) > 1 + \varepsilon \,\Big|\, \min_\ell N_{n-1,\ell}(w) > 0 \right) \le e^{-\frac{\varepsilon^2 \mathrm{L}}{4+2\varepsilon}}.$$

We apply this with $\varepsilon = \sqrt{6\ln(n^2/\delta)/\mathrm{L}}$. Since $\varepsilon \le 1$ the RHS is $\le \delta/n^2$. We deduce that for all $w \in A^*$

$$\begin{aligned}&\mathbb{P}\left(\left\| \left\{ \frac{d_\ell(\hat{p}_{n,\ell}(\cdot|w), \overline{p}_{n,\ell}(\cdot|w)}{\hat{\mathsf{cr}}_\ell(w)} \right\}_{\ell=1}^{\mathrm{L}} \right\|_{\mathrm{L},2} > 1 \,\Big|\, \min_{1 \le \ell \le \mathrm{L}} N_{n-1,\ell}(w) > 0 \right) \\ &\quad \le \frac{\delta}{n^2} \prod_{\ell=1}^{\mathrm{L}} \mathbb{P}\big(N_{n-1,\ell}(w) > 0\big).\end{aligned}$$

The rest of the proof follows the argument for the "General case." □

APPENDIX B: PROOF OF THEOREM 2

The proof of Theorem 2 follows from the oracle inequality in Theorem 1 restricted to $\mathcal{T}(h, \pi_*)$ and properly replacing the empirical confidence radii with the population confidence radii. Lemma 6 stated below (proven in the supplementary material [4]) establishes that the frequencies $N_{n-1,\ell}(w)$ are close to $\pi_\ell(w)n$ for typical trees provided a sample size condition holds. In turn, Lemma 7 below allows one to switch from empirical to population confidence radii, at the price of a small multiplicative constant.

In what follows, we use

$$\beta^{-1}(x) \equiv \min\{b \in \mathbb{N} : \forall b' \geq b, \beta(b') \leq x\} \qquad (x \in (0,1)).$$

LEMMA 6. *Let $X = (X_k)_{k\in\mathbb{Z}}$ be a stationary and β-mixing process over alphabet A with mixing rate function $\beta(\cdot)$. Consider a nonempty finite set $S \subset A^*$ and define*

$$h_S \equiv \max_{w\in S} |w|, \qquad \pi_S \equiv \min_{w\in S} \pi(w) \qquad \text{where } \pi(w) \equiv \mathbb{P}(X_{-|w|}^{-1} = w).$$

Let $\xi > 0$, $\delta_0 \in (0, 1/e)$ and $n \in \mathbb{N}$ satisfy

$$n \geq 2\left\{\left\lceil \frac{10h_S}{\xi} \right\rceil \vee \beta^{-1}\left(\frac{\xi\pi_S\delta_0}{24}\right)\right\} \times \left\{1 + \frac{300}{\xi^2\pi_S} \ln\left(\frac{12|S|}{\delta_0}\right)\right\},$$

then the random variables

$$N_n(w) \equiv \left|\{|w| \leq j \leq n : X_{j-|w|+1}^{j} = w\}\right|, \qquad w \in S,$$

satisfy

$$\mathbb{P}\left(\forall w \in S, 1 - \xi \leq \frac{N_n(w)}{\pi(w)n} \leq 1 + \xi\right) \geq 1 - \delta_0.$$

LEMMA 7. *Assume $X(1), \ldots, X(\mathrm{L})$ satisfy Assumptions 1 through 3, and the sample size n obeys*

$$n \geq 2\max\left\{40h, \left\lceil \frac{48\Gamma\mathrm{L}}{\pi_*\delta_0} \right\rceil^{1/\theta}\right\} \times \left\{1 + \frac{1200}{\pi_*} \log\left(\frac{24(h+1)}{\delta_0\pi_*}\right)\right\}.$$

Let

$$\mathrm{Typ}_r \equiv \bigcap_{\tilde{T}\in\mathcal{T}(h,\pi_*)} \bigcap_{w \text{ leaf of } \tilde{T}} \left\{\frac{\|\hat{\mathsf{cr}}(w)\|_{\mathrm{L},r}}{\|\overline{\mathsf{cr}}(w)\|_{\mathrm{L},r}} \leq \sqrt{2}\right\}.$$

Then $\mathbb{P}(\mathrm{Typ}_r) \geq 1 - \delta_0$.

PROOF. Define a set S consisting of all $w \in A^*$ of length $|w| \le h$ and $\min_\ell \pi_\ell(w) \ge \pi_*$. This set contains all leaves of trees $T \in \mathcal{T}(h, \pi_*)$, and it is clear from the definitions of confidence radii that

$$E = \bigcap_{\ell=1}^{\mathrm{L}} E_\ell \qquad \text{with } E_\ell \equiv \left\{ \forall w \in S : \frac{1}{2} \le \frac{N_{n-1}(w)}{n\pi_\ell(w)} \le \frac{3}{2} \right\}$$

is contained in Typ_r. We will apply the previous lemma to prove $\mathbb{P}(E_\ell) \ge 1 - \delta_0/\mathrm{L}$, which implies $\mathbb{P}(E) \ge 1 - \delta_0$ and completes the proof. We have processes $X(1), \dots, X(\mathrm{L})$ as in Definition 3, and choose parameters $n \ge 9$, $\xi = 1/2$, $\delta_0 = \delta_0/\mathrm{L}$. The mixing rate function $\beta(b) = \Gamma b^{-\theta}$ is the same for all processes. To obtain a bound on $|S|$, we note that

$$|S \cap A_{-k}^{-1}|\pi_* \le \sum_{w \in S \cap A_{-k}^{-1}} \mathbb{P}\big(X_{-k}^{-1}(1) = w\big) \le \sum_{w \in S \cap A_{-k}^{-1}} \pi_1(w) \le 1,$$

so

$$|S| \le \sum_{k=0}^{h} |S \cap A_{-k}^{-1}| \le \frac{h+1}{\pi_*}.$$

Thus, we see that, in order to apply Lemma 6 to $X(\ell)$, we need the condition

$$n \ge 2 \max\left\{ 40h, \left\lceil \frac{48\Gamma \mathrm{L}}{\pi_* \delta_0} \right\rceil^{1/\theta} \right\} \times \left\{ 1 + \frac{1200}{\pi_*} \log\left(\frac{24(h+1)}{\delta_0 \pi_*} \right) \right\},$$

which is precisely the assumption in the present lemma. This implies that Lemma 6 is indeed applicable, and we deduce $\mathbb{P}(E_\ell) \ge 1 - \delta_0/\mathrm{L}$, as desired. □

Acknowledgements. The authors would like to thank Victor Chernozhukov, Antonio Galves and Matthieu Lerasle for various discussions and to Whitney K. Newey for suggesting the dynamic choice model application. This work is part of USP project "Mathematics, computation, language and the brain."

SUPPLEMENTARY MATERIAL

Supplement to "Approximate group context tree" (DOI: 10.1214/16-AOS1455SUPP; .pdf). We provide additional discussion on the oracle context tree, omitted proofs from Section 5, a compendium of Martingale results, minimax rates for chain with infinite connections, and simulation results.

REFERENCES

[1] AGUIRREGABIRIA, V. and MIRA, P. (2010). Dynamic discrete choice structural models: A survey. *J. Econometrics* **156** 38–67. MR2609919

[2] ARELLANO, M. and HONORÉ, B. H. (2001). Panel data models: Some recent developments. *Handb. Econom.* **5** 3229–3296.

[3] BEJERANO, G. (2004). Algorithms for variable length Markov chain modeling. *Bioinformatics* **20** 788–789.

[4] BELLONI, A. and OLIVEIRA, R. I. (2016). Supplement to "Approximate group context tree." DOI:10.1214/16-AOS1455SUPP.

[5] BERTSEKAS, D. P. (1987). *Dynamic Programming*: *Deterministic and Stochastic Models*. Prentice Hall, Englewood Cliffs, NJ. MR0896902

[6] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[7] BROWNING, M. and CARRO, J. M. (2010). Heterogeneity in dynamic discrete choice models. *Econom. J.* **13** 1–39. MR2656540

[8] BROWNING, M. and CARRO, J. M. (2014). Dynamic binary outcome models with maximal heterogeneity. *J. Econometrics* **178** 805–823. MR3144684

[9] BÜHLMANN, P. (1999). Efficient and adaptive post-model-selection estimators. *J. Statist. Plann. Inference* **79** 1–9. MR1704215

[10] BÜHLMANN, P. (2000). Model selection for variable length Markov chains and tuning the context algorithm. *Ann. Inst. Statist. Math.* **52** 287–315. MR1763564

[11] BÜHLMANN, P. and WYNER, A. J. (1999). Variable length Markov chains. *Ann. Statist.* **27** 480–513. MR1714720

[12] CHERNOZHUKOV, V., FERNANDEZ-VAL, I., HAHN, J. and NEWEY, W. (2009). Identification and estimation of marginal effects in nonlinear panel models. Available at arXiv:0904.1990.

[13] CSISZÁR, I. and SHIELDS, P. C. (1996). Redundancy rates for renewal and other processes. *IEEE Trans. Inform. Theory* **42** 2065–2072.

[14] CSISZÁR, I. and TALATA, Z. (2006). Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory* **52** 1007–1016. MR2238067

[15] FARIAS, V. F., MOALLEMI, C. C., VAN ROY, B. and WEISSMAN, T. (2010). Universal reinforcement learning. *IEEE Trans. Inform. Theory* **56** 2441–2454. MR2729794

[16] FERRARI, F. and WYNER, A. (2003). Estimation of general stationary processes by variable length Markov chains. *Scand. J. Statist.* **30** 459–480. MR2002222

[17] GALVES, A., GALVES, C., GARCÍA, J. E., GARCIA, N. L. and LEONARDI, F. (2012). Context tree selection and linguistic rhythm retrieval from written texts. *Ann. Appl. Stat.* **6** 186–209. MR2951534

[18] GARIVIER, A. (2006). Redundancy of the context-tree weighting method on renewal and Markov renewal processes. *IEEE Trans. Inform. Theory* **52** 5579–5586. MR2300720

[19] GARIVIER, A. and LEONARDI, F. (2011). Context tree selection: A unifying view. *Stochastic Process. Appl.* **121** 2488–2506. MR2832411

[20] LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.* **35** 459–470. MR1091202

[21] LOUNICI, K., PONTIL, M., TSYBAKOV, A. B. and VAN DE GEER, S. (2010). Taking advantage of sparsity in multi-task learning. In *Proc. Computational Learning Theory Conference* (*COLT* 2009).

[22] OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.* **39** 1–47. MR2797839

[23] PUTERMAN, M. L. (1994). *Markov Decision Processes*: *Discrete Stochastic Dynamic Programming*. Wiley, New York. MR1270015

[24] RISSANEN, J. (1983). A universal data compression system. *IEEE Trans. Inform. Theory* **29** 656–664. MR0730903

[25] ROSS, S. (1983). *Introduction to Stochastic Dynamic Programming*. Academic Press, New York. MR0749232

[26] TALATA, Z. and DUNCAN, T. (2009). Unrestricted bic context tree estimation for not necessarily finite memory processes. In 2009 *IEEE International Symposium on Information Theory* 724–728.

[27] VERT, J.-P. (2001). Adaptive context trees and text clustering. *IEEE Trans. Inform. Theory* **47** 1884–1901. MR1842525

[28] WILLEMS, F. M. J., SHTARKOV, Y. M. and TJALKENS, T. J. (1995). The context-tree weighting method: Basic properties. *IEEE Trans. Inform. Theory* **41** 653–664.

[29] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

DUKE UNIVERSITY
100 FUQUA DRIVE
DURHAM, NORTH CAROLINA 27708
USA
E-MAIL: abn5@duke.edu

IMPA
ESTRADA DONA CASTORINA 110
RIO DE JANEIRO
BRAZIL 22460-320
E-MAIL: rimfo@impa.br