

APPROXIMATE GROUP CONTEXT TREE: APPLICATIONS TO DYNAMIC PROGRAMMING AND DYNAMIC CHOICE MODELS*

BY ALEXANDRE BELLONI AND ROBERTO I. OLIVEIRA[†]

Duke University and IMPA

The paper considers a variable length Markov chain model associated with a group of stationary processes that share the same context tree but potentially different conditional probabilities. We propose a new model selection and estimation method, develop oracle inequalities and model selection properties for the estimator. These results also provide conditions under which the use of the group structure can lead to improvements in the overall estimation.

Our work is also motivated by two methodological applications: discrete stochastic dynamic programming and dynamic discrete choice models. We analyze the uniform estimation of the value function for dynamic programming and the uniform estimation of average dynamic marginal effects for dynamic discrete choice models accounting for possible imperfect model selection.

We also derive the typical behavior of our estimator when applied to polynomially β -mixing stochastic processes. For parametric models, we derive uniform rate of convergence for the estimation error of conditional probabilities and 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.

Finally, we investigate the empirical performance of the proposed method in simulations and we apply this approach to account for possible heterogeneity across different years in a linguistic application.

1. Introduction. The dependence structure of stationary processes are of fundamental importance to understand the main features of the corresponding dynamic models. However, in many applications, the dependence structure is unknown. Not surprisingly, the development of models and

*First Version: June 2010; Current Version: November 27, 2011.

[†]Supported by projects *Universal* and *Produtividade em Pesquisa* from CNPq, Brazil. This work is part of USP project “Mathematics, computation, language and the brain”.

AMS 2000 subject classifications: Primary 62M05, 62M09, 62G05; secondary 62P2060J10

Keywords and phrases: categorical time series, group context tree, dynamic discrete choice models, dynamic programming, model selection, VLMC

model selection procedures for the estimation of the dependence structure have attracted substantial attention.

Variable length Markov chains (VLMCs) have emerged as a prominent model class for stationary processes by allowing a flexible and parsimonious representation of the dependence structure. The estimation of the dependence structure and the associated conditional probability distributions have been made possible by exploiting the intrinsic hierarchical tree structure. Importantly, despite of the exponentially large dimensionality of the model selection problem, efficient computational methods have been successfully developed starting with the seminal work of [23] proposing the Context algorithm.

The tractability and applicability of this model attracted considerable interest in different fields like statistics, information theory and machine learning. Subsequently, several papers have proposed various estimators and developed their statistical properties. [8] studies consistency when the underlying model increases in dimensionality as the sample size increases. They also proposed and shown the validity of a bootstrap scheme based on fitted VLMCs. [15] considers 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. BIC Context Tree algorithm and its consistency properties have been considered in [12], [14], [17] and [27]. Redundancy rates were studied by [13] and [17]. Typicality results were established by [11]. Several other works contributed to this literature in various directions [6], [7], [18] and others.

In this work we consider a group of stationary processes over a discrete alphabet. These stationary processes share the same dependence structure but possibly different conditional probability distributions across groups. We refer to this model as *group context tree* alluding to the recent literature on group lasso [21, 28]. As in the case of group lasso, by combining different processes which possess the same dependence structure we hope to improve the overall estimation. Interestingly, several data sets obey this model in many applications: in linguistic texts from different issues are used [16], in biology genetic sequences from different subjects [26], and others.

We propose a new estimator for model selection and estimation of conditional probabilities for a group context tree model which is based on confidence intervals. Intuitively, the length of a valid confidence interval is closely related to the variance in the estimation of a given group context tree, and therefore can be used as a regularization penalty. The bias of a model is related to the continuity rates of the transition probabilities, that is to say, to their approximability by Markovian models of finite order [19]. Bias and

variance are automatically balanced by a simple tree-pruning procedure that is reminiscent of Rissanen’s original Context algorithm.

We study several statistical properties of these estimators. We develop oracle inequalities, model selection properties, and typicality results. The analysis derives finite-sample bounds for the estimation errors in various norms. We show that the group context tree model can lead to improvements on the estimation when compared to the single-process case. Moreover, two salient features of our results seem to be new even in the single-process case:

- (i) The estimated transition probabilities are uniformly close to their true values whenever the latter are continuous functions of the infinite past. This turns out to be important in the applications discussed below.
- (ii) We obtain finite-sample bounds that account for the possible misspecification of the estimated model in a transparent way.

In addition, two new methodological applications motivated our investigation of the group context tree model. The first is the discrete stochastic dynamic programming problem in operations research [22, 24, 25]. In this problem a decision maker chooses among actions that yield some instantaneous reward and impact the system’s transition to the next state. That is, the transition probability distributions between states depend on the history of states and on the choice of action. Thus we have different processes for each action. Also, computational methods used in dynamic programming rely on having the same history of states across actions. The second methodological application is the dynamic discrete choice model in economics [1, 4]. This model consists of many different agents making choices over time. In order to account for heterogeneous agents, it is of interest to allow agents to have different conditional probability distributions at the same context but it is reasonable to expect that agents rely on the same context tree.

We also perform the analysis of the impact of possible misspecification and estimation errors from the group context tree in our two motivating applications. In these applications, the objects of main interest are not the conditional probabilities but rather functionals of them. In the discrete stochastic dynamic programming problem the object of interest is the value function over the history of states, which is defined as a fixed point of a suitable operator that depends on the transition probabilities. We derive uniform error bounds for the ℓ_∞ -norm between the fixed point computed based on probability estimates of the probabilities and the true value function. In discrete dynamic choice models several statistics associated with the consumer behavior of the agents are of interest. We focus our attention on the estimation of the average marginal dynamic effects. We derive uniform bounds on the

rate of convergence for the estimates of all the average marginal dynamic effect accounting for the model selection and possible misspecification.

Lastly, we discuss the typical behavior of our estimator when applied to polynomially β -mixing stochastic processes. Based on these results, two particular cases are worked out in detail. For parametric models, we derive uniform rate of convergence for the estimation error of conditional probabilities and perfect model selection results, essentially improving upon the assumptions of [8]. 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.

The paper is organized as follows. Section 2 formally introduces the approximate group context tree model and the proposed estimator. Section 3 contains the statistical analysis of the estimator. Section 5 is devoted to the two motivating applications. New typicality results for β -mixing processes are developed in Section 6. Simulations illustrating the performance of the estimator are presented in Section 7. Section 8 applies the group context tree model to understand the difference between the rhythmic features in European and Brazilian Portuguese accounting for heterogeneity. Finally, proofs are deferred to the Appendices.

2. Approximate group context tree. Let us introduce the notation and definitions. Let A denote a finite alphabet, A_{-k}^{-1} 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 $1 \leq k \leq |w|$, w_{-k}^{-1} is the suffix of w with length k . A subset $\tilde{T} \subset A^*$ is a *tree* if the empty string $e \in \tilde{T}$ and for all $w = w_{-|w|} \dots w_{-1} \in \tilde{T} \setminus \{e\}$ its suffix $w_{-|w|+1}^{-1} = w_{-|w|+1} \dots w_{-1}$, called the *parent* of w and denoted by $\text{par}(w)$, is also in \tilde{T} . 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 $x = \dots x_{-3}x_{-2}x_{-1} \in A_{-\infty}^{-1}$ a (possibly infinite) number

$$K_{\tilde{T}}(x) \equiv \sup\{k \in \mathbb{N} : x_{-k}^{-1} \in \tilde{T}\}.$$

We will write $\tilde{T}(x) = x_{-K_{\tilde{T}}(x)}^{-1}$. The strings of the form $\tilde{T}(x)$ where x ranges over $A_{-\infty}^{-1}$ will be called the *terminal nodes* of \tilde{T} .

REMARK 1 (Complete context trees). *It is sometimes convenient to work with complete trees, i.e. trees \tilde{T} where all non-leaf nodes have exactly $|A|$*

children. For a complete tree we have that if $x, y \in A_{-\infty}^{-1}$

$$y_{-K_{\tilde{T}}(x)}^{-1} = x_{-K_{\tilde{T}}(x)}^{-1} \Rightarrow \tilde{T}(x) = \tilde{T}(y).$$

It is not hard to show that the same result will not hold for incomplete trees; consider for instance $A = \{0, 1\}$, $\tilde{T} = \{e, 0, 1, 00\}$, $x = \dots 0010$, $y = \dots 0000$, $\tilde{T}(x) = 0$, and $\tilde{T}(y) = 00$.

In this paper a pair (\tilde{T}, \tilde{p}) will always 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 . The set of probability distributions over A will be denoted by Δ^A . A stationary ergodic process $X \equiv (X_n)_{n \in \mathbb{Z}}$ will be said to be *compatible* with (\tilde{T}, \tilde{p}) if:

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

If X is compatible with (\tilde{T}, \tilde{p}) , we say that \tilde{T} is a context tree for X . We will also sometimes write $p(a \mid x)$ instead of $p(a \mid \tilde{T}(x))$. There always exists a minimal complete context tree for X and we will always implicitly refer to that tree.

Finally, 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 χ_E , and for $q \geq 1$ the $\|\cdot\|_{L,q}$ -norm of a vector $v \in \mathbb{R}^L$ is defined as

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

2.1. Approximate group context tree model and oracle estimator. In an exact group context tree model we have L stationary processes, a context tree T^* , and probability distributions p_ℓ , $\ell = 1, \dots, L$, such that the ℓ th process is compatible with (T^*, p_ℓ) for $\ell = 1, \dots, L$. Note that T^* is possibly infinite so that this is not a restriction.

However, for a given sample, T^* might be too long to be efficiently estimated. Thus, in some cases, it is possible that a smaller tree, that is slightly misspecified, lead to much more efficient estimates for the conditional probabilities than T^* would due to the large variance. This motivates us to consider an oracle context tree that balances bias and variance instead of T^* as our main goal for estimation.

In order to define an approximate context tree model consider a metric $d_\ell : \Delta^A \times \Delta^A \rightarrow [0, 1]$ for each process, $\ell = 1, \dots, L$. For notational convenience we denote for $z, z' \in (A^*)^L$, $p(\cdot|z) = (p_1(\cdot|z(1)), \dots, p_L(\cdot|z(L)))$,

$d = (d_1, \dots, d_L)$, and

$$d(p(\cdot|z), \tilde{p}(\cdot|z')) = (d_1(p_1(\cdot|z(1)), \tilde{p}_1(\cdot|z'(1))), \dots, d_L(p_L(\cdot|z(L)), \tilde{p}_L(\cdot|z'(L)))).$$

By abuse of notation, we also apply the definitions above for $z \in A^*$ simply meaning $z = z(1) = \dots = z(L)$.

For every $\ell = 1, \dots, L$, we observe a sample $X_1^n(\ell) \equiv (X_1(\ell), \dots, X_n(\ell))$ from X . Given $w \in A^*$ and $k \geq |w| + 1$, we let $N_{k,\ell}(w)$ denote the number of indices i , $|w| \leq i \leq k$, with $X_{i-|w|}^i(\ell) = w$, that is, the number of occurrences of w in $X_1^k(\ell)$. For notational convenience we assume the length of the sample for each group is the same but the analysis does not rely on that.

Based on the sample, define the ‘‘oracle conditional probability’’ given a context w , $a \in A$, $\ell = 1, \dots, L$ as

$$\bar{p}_{n,\ell}(a|w) \equiv \frac{1}{N_{n-1,\ell}(w)} \sum_{i=|w|+1}^n \chi_{\{X_{i-|w|}^{i-1}(\ell)=w\}} p_\ell(a|X_{-\infty}^{i-1}(\ell)),$$

if $\min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) > 0$, and we define $\bar{p}_{n,\ell}(a|w) \equiv 1/|A|$ otherwise.

The conditional probability distribution $\bar{p}_{n,\ell}(\cdot|w)$ will play the role of an oracle estimate for the conditional probability $p_\ell(\cdot|w)$ which is adapted to the given sample. Thus \bar{p}_n is an intermediate step in the estimation of our ultimate goal p . Indeed, under mild regularity conditions it follows that $\bar{p}_{n,\ell}(a|X_{-k}^{-1}(\ell))$ converges to $p_\ell(a|X_{-\infty}^{-1}(\ell))$ as k and n grow at appropriate rates. For each context $w \in A^*$, we denote the approximation error of using $\bar{p}_n(\cdot|w)$ as an approximation for the underlying conditional probabilities as

$$(2.1) \quad c_w := \sup_{z \in (A_{-\infty}^{-1})^L} \|d(p(\cdot|z), \bar{p}_n(\cdot|w))\|_{L,k} : z(\ell) \succeq w, 1 \leq \ell \leq L.$$

Given a ‘‘confidence radius’’

$$\text{cf}(w) = (\text{cf}_1(w), \dots, \text{cf}_L(w)) \quad \text{for each } w \in A^*,$$

the context tree for the approximate model solves the following oracle problem for some $k \geq 1$ and $r \geq 1$:

$$(2.2) \quad \min_{\tilde{T}} \sup_{x \in A_{-\infty}^{-1}} c_{\tilde{T}(x)} + \left\| \text{cf}(\tilde{T}(x)) \right\|_{L,r}$$

where the minimum is over all finite complete trees. The choice of k and r typically depends on the final purpose of using the model (see Section 5 for examples).

The context tree T that solves the oracle problem (2.2) balances the bias of a misspecified model and the variance associated with its estimation measured as a function of the confidence radius. We discuss a suitable data-driven choice for cf in Section 4. For now, we just note that in general cf should depend on the metrics d_ℓ , $\ell = 1, \dots, L$, and we assume that $\text{cf}(w)$ is componentwise increasing in w , that is, $\text{cf}_\ell(w) \leq \text{cf}_\ell(w')$ if $w' \succeq w$. For convenience we also assume that $0 \leq \text{cf}_\ell(w) \leq 1$ for all $w \in A^*$.

REMARK 2 (On the oracle problem, non-uniqueness). *The oracle problem (2.2) might have multiple solutions. Although the results derived here allow for any such solution to be considered, the oracle further selects a context tree by fixing the paths which achieve the optimal value of the oracle's objective function and further minimization of the criterion function over the remaining paths.*

REMARK 3 (On the oracle problem, approximation error). *Under mild conditions on the processes the oracle can adjust the length of the contexts in the oracle tree to make the approximation error $c_{T(x)}$ to be (at most) of the same order as the regularization term, namely, there is a constant K such that uniformly in $x \in A_{-\infty}^{-1}$ we have*

$$(2.3) \quad c_{T(x)} \leq K \|\text{cf}(T(x))\|_{L,r}.$$

2.2. *Model selection and estimation of conditional probabilities.* Next we discuss the model selection method which leads to the estimation of the conditional probabilities. The method relies on the length of confidence intervals associated with each suffix as a regularization term. These regularization terms are used to decide if a particular node should be pruned from the tree. After selecting a context tree, probabilities compatible with it are computed. Next we describe in detail the procedure.

For each $w \in A^*$ with $\min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) > 0$, we define:

$$\hat{p}_{n,\ell}(a|w) \equiv \frac{N_{n,\ell}(wa)}{N_{n-1,\ell}(w)}, \quad \text{for } a \in A, \ell = 1, \dots, L$$

as an empirical estimate for the conditional probability distribution $p(a|w)$. For definiteness we set $\hat{p}_{n,\ell}(a|w) \equiv 1/|A|$ if $\min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) = 0$.

Let E_n be the suffix tree that contains every string $w \in A^*$ such that

$\min_{\ell=1,\dots,L} N_{n-1,\ell}(w) > 0$. For a fixed constant $c > 1$ define:

$$(2.4) \quad \text{CanRmv}(w) \equiv \begin{cases} 1, & \text{if for all } w', w'' \in E_n \text{ with } w \preceq w', \text{par}(w) \preceq w'' \\ & \|d(\hat{p}_n(\cdot|w'), \hat{p}_n(\cdot|w''))\|_{L,k} \leq c\|\text{cf}(w')\|_{L,r} + c\|\text{cf}(w'')\|_{L,r}; \\ 0, & \text{otherwise.} \end{cases}$$

The constant c is similar in spirit to the slack parameter used in [3]; we recommend $c = 1.01$ in practice. Intuitively, $\text{CanRmv}(w) = 1$ only if removing w from the tree will not make a significant impact in the estimation of conditional probabilities relative to the noise in their estimation. We prune E_n into a tree \hat{T}_n via the pruning algorithm in Figure 1 to obtain our estimate for the oracle context tree T defined as the solution of (2.2).

Procedure PruneTree

```

1:  $\hat{T}_n \leftarrow E_n$ . ▷ In the beginning,  $\hat{T}_n$  contains all visible strings
2: for each node  $w$  of  $E_n$  do
3:    $\text{exam}(w) \leftarrow 0$  ▷ All nodes start out unexamined
4: end for
5: while  $\exists$  leaf  $w \in \hat{T}_n$  with  $\text{exam}(w) = 0$  do ▷ While there are unexamined leaves
6:   if  $\text{CanRmv}(w) = 1$  then ▷ If  $w$  can be removed, remove it.
7:      $\hat{T}_n \leftarrow \hat{T}_n \setminus \{w\}$ 
8:   end if
9:    $\text{exam}(w) \leftarrow 1$  ▷  $w$  has been examined
10: end while
11: return  $\hat{T}_n$ 

```

FIG 1. *The pruning algorithm.*

After selecting the model \hat{T}_n we proceed to estimate the conditional probability distributions. Given $x \in A_{-\infty}^{-1}$, we will assign a conditional probability $\hat{P}_n(\cdot|x)$ which is compatible with \hat{T}_n , as follows:

$$\hat{P}_n(\cdot|x) \equiv \hat{p}_n(\cdot|\hat{T}_n(x)) \equiv \hat{p}_n\left(\cdot|x_{-K_{\hat{T}_n}(x)}^{-1}\right).$$

REMARK 4 (Complete context trees, continued). *In general, \hat{T}_n will not be complete in the sense of Remark 1. Completeness can always be achieved by adding leaves nodes to non-leaves nodes of the tree created by the algorithm has finished. In that case, the conditional probabilities for the added leaves are set to the conditional probabilities of their corresponding parent which was not pruned by the algorithm.*

REMARK 5 (Variations of CanRmv). *All results we will establish in the following sections would remain true if in the definition of $\text{CanRmv}(w)$ in*

(2.4) we set $w'' \in \mathcal{W}$ where $\text{par}(w) \in \mathcal{W} \subseteq \{z \in E_n : z \succeq \text{par}(w)\}$. 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.

REMARK 6 (Computational efficiency). *The algorithm can be implemented efficiently, i.e. in polynomial time with respect to the parameters L and n of the data. Observe that $\text{CanRmv}(w)$ can be computed efficiently from the list of values:*

$$\text{List}(w) \equiv \{(\hat{p}_n(\cdot|w'), \text{cf}(w')) : w' \in E_n, w' \succeq w\}$$

and the corresponding list for $\text{par}(w)$. Since $\text{CanRmv}(w)$ 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 $\text{List}(w)$. This can be achieved as follows.

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

$$\text{List}(\text{par}(w)) \leftarrow \text{List}(\text{par}(w)) \cup \left(\bigcup_{w' \in \hat{T}_n : \text{par}(w') = \text{par}(w)} \text{List}(w') \right).$$

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

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

3. Analysis. In this section we derive our main theoretical results on the performance of the estimates proposed in Section 2. We start by characterizing an event which ensures that the pruning and estimation procedures behave properly. Then we proceed to establish model selection results and oracle inequalities for the estimation of the conditional probabilities when using the algorithm described in the previous section.

3.1. *Regularization event.* The sources of the estimation error are the deviation of $\hat{p}_{n,\ell}(\cdot|w)$ from the oracle conditional probabilities $\bar{p}_{n,\ell}(\cdot|w)$ and the approximation error of using the latter instead of p_ℓ . Under this time series framework, the key issues are the need to simultaneously estimate these probabilities for a number of suffixes $w \in A^*$ that is substantially larger than the sample size n , and to select one model (i.e. a context tree) among the exponentially many possible models.

The underlying idea will be to rely on some regularization which would prune from the estimated tree suffixes that contain a large noise relative to the explanatory power they could bring to the model. Hopefully, this would reduce the number of suffixes in the estimated context tree leading to consistent estimates of the conditional probabilities. Our main insight is to rely on the length of a confidence interval for $\bar{p}_{n,\ell}(\cdot|w)$ as the main ingredient for the regularization penalty.

The regularization penalty in the pruning algorithm consists of a confidence radius $\text{cf}_\ell(w)$ for each suffix $w \in A^*$ and process $\ell = 1, \dots, L$, so that with high probability the following event occurs for a prescribed $m \geq 1$:

$$\text{Good}_m \equiv \bigcap_{w \in A^*} \left\{ \left\| \left\{ \frac{d_\ell(\bar{p}_{n,\ell}(\cdot|w), \hat{p}_{n,\ell}(\cdot|w))}{\text{cf}_\ell(w)} \right\}_{\ell=1}^L \right\|_{L,m} \leq 1, \text{ if } \min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) > 0 \right\}.$$

By the definition of $\|\cdot\|_{L,m}$, if $d_\ell(\bar{p}_{n,\ell}(\cdot|w), \hat{p}_{n,\ell}(\cdot|w)) \leq \text{cf}_\ell(w)$ for every $\ell = 1, \dots, L$, the event Good_m occurs but it is also likely to occur if this condition holds for most values of ℓ . For efficiency reasons we will aim to choose the smallest confidence radius such that

$$\mathbb{P}(\text{Good}_m) \geq 1 - \delta$$

where $1 - \delta$ is our desired confidence level. In Section 4 we will propose and analyze data-driven choices for the confidence radius so that Good_m occurs with high probability.

The following condition summarized our setup.

CONDITION 1 (AGCT). *We have data $\{X_1^n(\ell) \in A^n, \ell = 1, \dots, L\}$ that for each n obey the model in Section 2, with T defined by (2.2) where $\text{cf}_\ell(w) \leq 1$ is componentwise increasing in $w \in A^*$, for every $1 \leq \ell \leq L$. The approximation error $c_{T(x)}$ is defined as in (2.1). The positive (extended) integers k, r and m satisfy $k \leq m$ and $r \geq km/(m - k)$ (equivalently $k \leq r, m \geq kr/(r - k)$), or $k \leq r = m = \infty$.*

The event Good_m will imply many desirable properties of the estimators including oracle inequalities. The following result relates the criterion function with the regularization term used in the oracle under the event Good_m .

THEOREM 1. *Suppose Condition AGCT holds and that the event Good_m occurs. Then, for all $w \in A^*$, we have*

$$\|d(\hat{p}_n(\cdot|w), \bar{p}_n(\cdot|w))\|_{L,k} \leq \|\text{cf}(w)\|_{L,r}.$$

Thus, if Good_m occurs, up to the regularization, $\hat{p}_n(\cdot|w)$ can be used as a good approximation for the oracle condition probabilities $\bar{p}_n(\cdot|w)$ uniformly over $w \in A^*$. An immediate corollary for the estimation of the true conditional probabilities follows by triangle inequality.

COROLLARY 1. *Suppose Condition AGCT holds and that the event Good_m occurs. Then we have for all $x \in A_{-\infty}^{-1}$*

$$\|d(\hat{p}_n(\cdot|T(x)), p(\cdot|x))\|_{L,k} \leq c_{T(x)} + \|\text{cf}(T(x))\|_{L,r}.$$

In words, this corollary establishes that $\hat{p}_n(\cdot|T(x))$ is also a good approximation to $p(\cdot|x)$ itself if the event Good_m occurs.

Up to this point, the parameter $c > 1$ used in the algorithm proposed in Section 2 has not played a role. The role of c will be to allow us to extend the nice properties established above for the oracle context tree T to the estimated context tree \hat{T}_n accounting for the model selection which can lead to a misspecified estimated model.

3.2. Model selection results. Next we derive properties of the estimated context tree \hat{T}_n relatively to the oracle context tree T and the (possibly infinite) compatible context tree T^* of the processes. Recall that as mentioned in Section 2, we have $T \subseteq T^*$. The next result addresses a similar question for the estimated context tree \hat{T}_n .

THEOREM 2. *Suppose Condition AGCT holds, and that the event Good_m occurs. Then we have $\hat{T}_n \subseteq T^*$.*

In particular, Theorem 2 shows that by a suitable choice of the confidence radius cf we have that the estimated context tree does not overestimate a compatible context tree T^* with high probability.

As discussed before, a compatible context tree might be too long and might not lead to the most efficient estimation of the conditional probabilities. This is the underlying motivation for the oracle context tree T . It balances bias and variance to achieve a good estimation performance. Thus a more interesting question is how does the estimated context tree \hat{T}_n compare with the oracle context tree T . Although some branches of \hat{T}_n can be longer than the corresponding branches of T , we can show that they cannot be too much longer when measured in the regularization function.

THEOREM 3. *Suppose Condition AGCT holds, T is a complete tree, and the event Good_m occurs. We have that for all $x \in A_{-\infty}^{-1}$*

$$\|\text{cf}(\hat{T}_n(x))\|_{L,r} \leq \max \left\{ \|\text{cf}(T(x))\|_{L,r}, \frac{2c_{T(x)}}{c-1} - \|\text{cf}(T(x))\|_{L,r} \right\}.$$

This result establishes an oracle inequality for the regularization term of the leaves of \widehat{T}_n . The slack parameter $c > 1$ combined with the Good_m event allows to restrict the $\|\cdot\|_{L,r}$ -size of \widehat{T}_n . In particular, in the typical case in which the oracle context tree yields $c_{T(x)} \lesssim \|\text{cf}(T(x))\|_{L,r}$, the inequality above achieves $\|\text{cf}(\widehat{T}_n(x))\|_{L,r} \lesssim \|\text{cf}(T(x))\|_{L,r}$.

It is possible to derive bounds on the actual length of $\widehat{T}_n(x)$ relative to $T(x)$ by relying on the following regularity condition.

CONDITION 2 (RL). *There is a $\kappa \in (0, 1)$ and integer $\bar{k} \geq 1$ such that*

$$\sup_{x \in A_{-\infty}^{-1,k}} \left\{ \frac{\|\text{cf}(x_{-k}^{-1})\|_{L,r}}{\|\text{cf}(x_{-k-\bar{k}}^{-1})\|_{L,r}} : |T(x)| \leq k \leq |T(x)| - \frac{\bar{k} \log \|\text{cf}(T(x))\|_{L,r}}{\log(1/\kappa)} \right\} \leq \kappa.$$

Condition RL is similar to the modulus of continuity between the regularization penalty and the length in a neighborhood of $T(x)$. It turns out that such condition is satisfied for many designs of interest. Under Condition RL, we can establish the following result.

THEOREM 4. *Suppose that Conditions AGCT and RL hold, and that the event Good_m occurs. Then for all $x \in A_{-\infty}^{-1}$ we have that*

$$|\widehat{T}_n(x)| \leq |T(x)| + \frac{\bar{k}}{\log(1/\kappa)} \max \left\{ 0, \log \left(\frac{2}{c-1} \frac{c_{T(x)}}{\|\text{cf}(T(x))\|_{L,r}} \right) \right\}.$$

It is interesting to consider the result of Theorem 4 when (2.3) holds. In this case we have

$$|\widehat{T}_n(x)| \leq |T(x)| + \frac{\bar{k}}{\log(1/\kappa)} \max \{0, \log K + \log(2/(c-1))\}.$$

Moreover, under mild conditions on the process, κ is bounded away from zero and \bar{k} is bounded above uniformly in n . Therefore, these regularity conditions imply that the length of $\widehat{T}_n(x)$ is not larger than the length of $T(x)$ plus a constant factor.

3.3. Oracle inequality for conditional probability estimation. Next we focus on the estimation of the conditional probabilities. In particular our goal is to derive uniform bounds over sequences $x \in A_{-\infty}^{-1}$ between the true conditional probability distributions $p(\cdot|x)$ and their estimate $\widehat{P}_n(\cdot|x) = \widehat{p}_n(\cdot|\widehat{T}_n(x))$. Our main result is the following oracle inequality.

THEOREM 5. *Suppose Condition AGCT holds and that the event Good_m occurs. Then for all $x \in A_{-\infty}^{-1}$ we have*

$$\begin{aligned} \left\| d(\widehat{P}_n(\cdot|x), \bar{p}_n(\cdot|T(x))) \right\|_{L,k} &\leq \max \left\{ (1+2c) \|\text{cf}(T(x))\|_{L,r}, \frac{c+1}{c-1} c_{T(x)} \right\} \text{ and} \\ \left\| d(\widehat{P}_n(\cdot|x), p(\cdot|x)) \right\|_{L,k} &\leq c_{T(x)} + \max \left\{ (1+2c) \|\text{cf}(T(x))\|_{L,r}, \frac{c+1}{c-1} c_{T(x)} \right\}. \end{aligned}$$

The combination of the event Good_m and the parameter $c > 1$ in the algorithm allows to derive finite-sample guarantees on the estimation performance of the conditional probabilities. Furthermore, in the typical case that $c_{T(x)} \lesssim \|\text{cf}(T(x))\|_{L,r}$, under the conditions of Theorem 5 we have

$$\left\| d(\widehat{P}_n(\cdot|x), p(\cdot|x)) \right\|_{L,k} \lesssim \|\text{cf}(T(x))\|_{L,r}$$

recovering the same performance (up to a constant factor) of the oracle estimator.

REMARK 7. *The results in Theorem 5 hold for any tree \widetilde{T} and not only the oracle tree T . In fact, by plugging in a compatible context tree T^* it yields*

$$\left\| d(\widehat{P}_n(\cdot|x), p(\cdot|x)) \right\|_{L,k} \leq (1+2c) \|\text{cf}(T^*(x))\|_{L,r}.$$

4. Data-driven choices for confidence radius. In this section we propose and analyze data-driven choices for setting the confidence radius $\text{cf}(w)$ for all $w \in A^*$. Intuitively, the confidence radius should majorate the deviations of $\widehat{p}_{n,\ell}$ from $\bar{p}_{n,\ell}$ which is the noise in our model selection problem. Large confidence radius overcome the noise easily but introduces a large bias. Small confidence radius would not allow for consistent estimation due to the large number of potential models.

By definition of the event Good_m , a proper choice of confidence radius cf will depend on the metrics d_ℓ and on the choice of m . We will consider in detail the (extreme) cases that are the relevant ones in the methodological applications in Section 5: $m = 2$ and $m = \infty$. Regarding the choice of metrics, here we will consider $d_\ell = \|\cdot\|_\infty$ for every $\ell = 1, \dots, L$, and $d_\ell = \|\cdot\|_1/2$ for every $\ell = 1, \dots, L$. An unifying approach to work with these metrics consists of choosing an appropriate family of sets $\mathcal{S} \subset 2^A$ to write them as

$$d_\ell(p_\ell, \tilde{p}_\ell) = \sup_{S \in \mathcal{S}} |p_\ell(S) - \tilde{p}_\ell(S)|$$

where $\|\cdot\|_\infty$ and corresponds to $\mathcal{S} = \{a \in A\}$ and $\|\cdot\|_1/2$ corresponds to $\mathcal{S} = 2^A$.

Each set $S \in \mathcal{S}$ induces a martingale so deviations between $\hat{p}_{n,\ell}(S|w)$ and $\bar{p}_{n,\ell}(S|w)$ can be controlled by martingale inequalities developed in Appendix E. Several other combinations can be derived directly from the presented analysis. Nonetheless, the cases covered here will allow us to compare interesting features of group context tree models relative to the traditional (single-process) context tree model.

In order to clearly communicate the main results we simplify the constants in the data-driven choices in the main text. In the appendix we state precise results that share the same asymptotic rates. For notational convenience define a function, for $w \in A^*$, $\epsilon \in (0, 1)$, $\ell = 1, \dots, L$:

$$c_\ell(w, \epsilon) := 2\sqrt{\frac{1}{N_{n-1,\ell}(w)}}\sqrt{\log(1/\epsilon) + 2\log(2 + 2\log N_{n-1,\ell}(w))}.$$

The value of ϵ is set to account for the model selection and the confidence level $1 - \delta$ of the event Good_m occurring. The $\log \log N_{n-1,\ell}(w)$ factor emerges from the time dependence in the data. In the traditional single-process case, $L = 1$, ϵ is chosen so that $\log 1/\epsilon = O(\log(n/\delta))$. This is similar to the rate in the case of Good_∞ for the group context tree.

THEOREM 6. *Let $\delta \in (0, 1)$, and for $w \in A^*$ with $\min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) > 0$, $\ell = 1, \dots, L$, let*

$$\text{cf}_\ell^I(w) := c_\ell\left(w, \frac{\delta}{n^2|\mathcal{S}|L}\right).$$

Then, setting the confidence radius to be $\text{cf}_\ell(w) = \min\{\text{cf}_\ell^I(w), 1\}$, with probability at least $1 - \delta$ the event Good_∞ occurs.

Thus, the rate for $\text{cf}_\ell^I(w)$ is $\sqrt{\log(nL/\delta)/N_{n-1,\ell}(w)}$ recovering the rate of a single group in the typical case that $\log L \lesssim \log n$. However, for other choices of m , it is possible to improve upon that rate by exploiting the context tree estimation across different groups. The following result addresses this question for $m = 2$.

THEOREM 7. *Suppose Condition AGCT holds. Let $\delta \in (0, 1)$, and assume that for some $\alpha < 3$ we have $\log^2(n^2L/\delta) \leq \alpha L \log \log n$, and for $w \in A^*$, $\ell = 1, \dots, L$, let*

$$\text{cf}_\ell^I(w) = c_\ell\left(w, \frac{\log \log n}{4|\mathcal{S}|\log^2(n^2L/\delta)}\right)\sqrt{(1 + 3\alpha/2)\left(1 + \frac{1}{\log n}\right)}.$$

Then, setting the confidence radius to be $\text{cf}_\ell(w) = \min\{\text{cf}_\ell^I(w), 1\}$, the event Good_2 occurs with probability at least $1 - \delta$.

In this case, if $\log^2(n/\delta) = o(L)$, the rate of $\text{cf}_\ell^I(w)$ 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 oracle model was known in advance.

4.1. *Improvement based on maximal variance.* The choices above do not explore the intrinsic variance within of the norm, namely

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

Although this factor does not affect the rate in general, in finite-sample it can yield an important factor to avoid over regularization. This is particularly evident in the case of $d_\ell = \|\cdot\|_\infty$ with $|A| > 2$.

Our second data-driven choice will account for that to generate strictly smaller confidence radius with the same guarantee. 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

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

When this is not the case, we can still the previous choice $\text{cf}_\ell^I(w)$. To concisely state the results regarding the maximum variance we define

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

THEOREM 8. *Suppose Condition AGCT holds, choose $\delta \in (0, 1)$, and let $m = 2$ or $m = \infty$. For $m = 2$ suppose further that assumptions of Theorem 7 hold. For $w \in A^*$, $\ell = 1, \dots, L$ set*

$$\text{cf}_\ell^{\tilde{\sigma}}(w) := \tilde{\sigma}_\ell(w) \text{cf}_\ell^I(w)$$

where $\text{cf}_\ell^I(w)$ is chosen as in Theorem 6 if $m = \infty$ and as in Theorem 7 if $m = 2$. Then, setting the confidence radius to be $\text{cf}_\ell(w) = \min\{\text{cf}_\ell^{\tilde{\sigma}}(w), 1\}$, we have that with probability at least $1 - \delta$ the event Good_m occurs.

By construction, it follows that $\text{cf}_\ell^{\tilde{\sigma}}(w) \leq \text{cf}_\ell^I(w)$ since $\tilde{\sigma}_\ell(w) \leq 1$. However, $\text{cf}_\ell^{\tilde{\sigma}}(w)$ might not be non-increasing in w . Nonetheless, the confidence radius

$\text{cf}^{\bar{\sigma}}$ can be majorated by the monotone confidence radius which still leads to an improvement over cf^I , namely

$$\text{cf}_\ell^*(w) = \max_{w' \preceq w} \text{cf}_\ell^{\bar{\sigma}}(w') \leq \max_{w' \preceq w} \text{cf}_\ell^I(w') = \text{cf}_\ell^I(w).$$

A side remark is that $\text{cf}_\ell^*(w)$ requires the estimation of $\bar{\sigma}_\ell(w)$. It follows that any such estimator will satisfy $\hat{\sigma}_\ell(w) \leq 1/2$ still achieving smaller confidence radius than cf^I . However, the estimates need to satisfy $\bar{\sigma}_\ell(w) \leq \hat{\sigma}_\ell(w)$ with high probability.

5. Methodological applications. In this section we develop two applications of the approximate group context tree model and estimation algorithms. In both cases the main object of interest is not the context tree nor the conditional probabilities but 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.

5.1. *Discrete stochastic dynamic programming.* Discrete stochastic dynamic programming 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 a A -valued random function $s(x_{-\infty}^t, u)$ which transition probabilities 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:

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

where $\beta < 1$ is the discount factor and $x s(x, u)$ is the concatenation of x with $s(x, u)$.

The success of this approach relies on avoiding to use the complete state space $A_{-\infty}^{-1}$. The selected state space needs to be rich enough to capture the main features of the transition function $s(\cdot, \cdot)$. However, in practice the transition probabilities between states need to be estimated, and algorithms to compute the value function can suffer from a curse of dimensionality if the state space is large.

Thus, our motivation to apply the methods described in the previous section is two fold. First, to create estimates for the transition probabilities.

Second, to create a data-driven manageable state space. This is exactly the case in which using the AGCT model can be more attractive than using a substantially larger compatible tree T^* . We advocate in favor of a small approximation error (that is 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

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

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

The following result derive guarantees on the estimation error between the value function V and its estimator \widehat{V} that solves the fixed point problem defined with the estimates of the conditional probabilities $\widehat{P}_{n,u}(\cdot | x)$.

LEMMA 1. *In the discrete stochastic dynamic programming problem described above, for $q \geq 1$ the estimator \widehat{V} satisfies*

$$\max_{x \in A_{-\infty}^{-1}} \frac{|\widehat{V}(x) - V(x)|}{\|V(x)\|_{\frac{q}{q-1}} \max_{u \in \mathcal{U}} \|\widehat{P}_{n,u}(\cdot | x) - p_u(\cdot | x)\|_q} \leq \frac{\beta}{1 - \beta}$$

where $q/(q-1) = \infty$ if $q = 1$.

The lemma above allows us to apply the results on the estimation of the conditional probabilities to the stochastic dynamic programming problem. Let the number of groups $L = |\mathcal{U}|$, $d_\ell = \|\cdot\|_1/2$ and $k = r = \infty$.

THEOREM 9. *In the discrete stochastic dynamic programming problem described above, by choosing \mathbf{cf} as in Theorem 6, we have that with probability at least $1 - \delta$ the estimator \widehat{V} of the value function satisfies*

$$\max_{x \in A_{-\infty}^{-1}} \frac{|\widehat{V}(x) - V(x)|}{\|\mathbf{cf}(T(x))\|_{L,\infty} \|V(x)\|_{\infty}} \leq \frac{2\beta}{1 - \beta} \left(1 + 2c + \frac{2c}{c-1} \sup_{x \in A_{-\infty}^{-1}} \frac{c_{T(x)}}{\|\mathbf{cf}(T(x))\|_{L,\infty}} \right)$$

$$\text{where } \|\mathbf{cf}(T(x))\|_{L,\infty} \lesssim \max_{\ell=1,\dots,L} \sqrt{\frac{\log(nL/\delta) + \log \log N_{n-1,\ell}(T(x)) + |A|}{N_{n-1,\ell}(T(x))}}.$$

Recall that under mild conditions the oracle balances bias and variance so that there is a constant K such that

$$\sup_{x \in A_{-\infty}^{-1}} \frac{c_{T(x)}}{\|\mathbf{cf}(T(x))\|_{L,r}} \leq K.$$

In this case, the rate of convergence is governed by the oracle regularization term $\|\mathbf{cf}(T(x))\|_{L,r}$ which goes to zero as the sample size increases.

5.2. *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, 4, 5, 9]. Typically, models pre-specify the Markovian structure of the process which is commonly assumed to be a Markov Chain 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 [4, 5].

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 $a \in A$, $x, y \in A_{-\infty}^{-1}$

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

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 [4].

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 ℓ th agent is

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

and the estimator for the average marginal dynamic effect is

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

This motivates the choice of $d_\ell = \|\cdot\|_\infty$, $k = 1$, and $r = m = 2$ in the AGCT model.

THEOREM 10. *In the dynamic discrete choice model described above, if the context tree and conditional probabilities are estimated with cf as in Theorem 7, we have that with probability at least $1 - 2\delta$ the estimator for*

the average marginal dynamic effect satisfies

$$\max_{\substack{a \in A, \\ x, y \in A_{-\infty}^{-1}}} \frac{|\widehat{\text{AVEm}}(a, x, y) - \text{AVEm}(a, x, y)|}{c_{T(x)} + c_{T(y)} + \|\text{cf}(T(x))\|_{L,2} + \|\text{cf}(T(y))\|_{L,2} + \sqrt{\frac{2 \log\left(\frac{|A| \cdot n^4}{4\delta}\right)}{L}} + \frac{2}{L}} \leq \frac{4c^2}{c-1}.$$

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

This uniform rate of convergence for the average marginal dynamic effect is governed by the rate of convergence of the conditional probabilities of the oracle estimator, and the number of different agents in the data. As mentioned in the previous section, in many models of interest there is a constant K such that

$$\sup_{z \in A_{-\infty}^{-1}} \frac{c_{T(z)}}{\|\text{cf}(T(z))\|_{L,2}} \leq K$$

so that the rate of convergence of the conditional probabilities is governed by the regularization terms and $\sqrt{\log n/L}$. Interestingly, the above result holds uniformly over all pairs $x, y \in A_{-\infty}^{-1}$. The cost to attain this uniform rate that accounts for the model selection and the size of \widehat{T}_n is the $\sqrt{\log n}$.

6. Bounds under primitive conditions. In this section we discuss finite-sample behavior of the AGCT estimator for families of processes satisfying certain conditions. One key ingredient will be to derive finite-sample bounds on the behavior of the confidence radius under appropriate mixing conditions. Our basic setup is summarized by the following assumption.

ASSUMPTION 1 (Basic Setup). $\{X(\ell)_{-\infty}^{+\infty}\}_{\ell=1}^L$ are stationary stochastic processes with values in a finite alphabet A . We are given $n \in \mathbb{N}$ with $n \geq 9$, $L \in \mathbb{N}$ is the number of groups and $\delta \in (0, 1)$ is a confidence parameter. We write $L = n^{\alpha_L}$, $\delta = n^{-\alpha_\delta}$ and

$$\alpha \equiv \alpha_L + \alpha_\delta > 0 \text{ as } 0 < \delta < 1.$$

We let $(\widehat{T}_n, \widehat{P}_n)$ denote the output of the AGCT estimator on input $\{X(\ell)_1^n\}_{\ell=1}^L$ with confidence level $1 - \delta$, a metric $d = d_S$, slack parameter $c > 1$ with the confidence interval described in Section 4.

In what follows, we will need additional notation in which $C_1, \dots, C_6 > 0$ are universal constants (i.e., they do not depend on the processes, on their mixing rates, or on the parameters n, L, δ). Given $w \in A^*$ and $k \geq 1$, define

$$\bar{c}_w \equiv \sup_{x, y \in (A_{-\infty}^{-1})^L} \|d(p(\cdot | x), p(\cdot | y))\|_{L,k} : x_{-|w|}^{-1}(\ell) = y_{-|w|}^{-1}(\ell) = w, 1 \leq \ell \leq L.$$

The approximation error \bar{c}_w is more primitive than c_w which depends on the particular sample. This new quantity will allow us to make statements based only on the properties of the processes. Nonetheless these approximation errors are closely related and they satisfy $c_w \leq \bar{c}_w \leq 2c_w$, for all $w \in A^*$. Hence, for the purpose of rates of convergence they are interchangeable. We also define steady state probabilities

$$\pi_\ell(w) \equiv \mathbb{P}\left(X(\ell)_{-|w|}^{-1} = w\right), \quad 1 \leq \ell \leq L.$$

For a complete finite tree $\tilde{T} \subset A^*$, let $h_{\tilde{T}}$ denote the height of \tilde{T} and

$$\pi_{\tilde{T}} \equiv \min_{1 \leq \ell \leq L} \min \left\{ \pi_\ell(w) : w \text{ leaf of } \tilde{T}, \pi_\ell(w) > 0 \right\} > 0.$$

In order to derive primitive rates of convergence, we will control the data-driven confidence radius $\text{cf}(w)$ which determines the size of the regularization term. We define non-empirical versions of the confidence radii $\bar{\text{cf}}(w)$ by replacing $N_{n-1,\ell}(w)$ with $n\pi_\ell(w)$ in the definition of $\text{cf}_\ell(w)$. Given a complete finite tree \tilde{T} and $r \geq 1$, the following typicality event plays a central role:

$$(6.1) \quad \text{Typ}_r(\tilde{T}) \equiv \left\{ \sup_{x \in A_{-\infty}^{-1}} \left| \frac{\|\text{cf}(\tilde{T}(x))\|_{L,r}}{\|\bar{\text{cf}}(\tilde{T}(x))\|_{L,r}} - 1 \right| \leq \frac{C_2}{\log n} \right\}.$$

6.1. *General results under β -mixing conditions.* We recall the definition of β -mixing.

DEFINITION 1. *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 $\beta : \mathbb{N} \rightarrow [0, 1]$ with $\lim_{b \in \mathbb{N}, b \rightarrow \infty} \beta(b) = 0$ and $\forall k \in \mathbb{Z}, s \in \mathbb{N}$:*

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

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

We will present general results under the assumption of polynomial β -mixing.

ASSUMPTION 2 (Beta Mixing). *Assumption 1 holds and there are constants $\Gamma > 0$ and $\gamma > 0$ such that the processes $X(1), \dots, X(L)$ are β -mixing with common rate function:*

$$\beta(b) \equiv \Gamma b^{-\gamma} \quad (b \in \mathbb{N}).$$

REMARK 8. *Most of the literature on context-tree-based estimation assumes ϕ -mixing or stronger conditions. ϕ -mixing consists of replacing the expectation in Definition 1 by an essential supremum, and can be shown for finite-order Markov chains and other examples. However, there are other natural examples (such as renewal processes) that are β -mixing but not ϕ -mixing. Ferrari and Wyner [15] make the alternative assumption of geometric α -mixing. We will provide a more detailed comparison after stating our results below; for now we only note that this assumption, as well as some of their other assumptions, are incomparable with ours.*

For $\delta_0 \in (0, 1)$ define the set of complete trees whose leaves are neither too rare nor too long relative to the β -mixing condition and the sample size n :

$$(6.2) \quad \mathcal{T}_{\delta_0} = \left\{ \begin{array}{l} \tilde{T} \subset A^* \\ \text{complete finite tree} \end{array} : \begin{array}{l} \pi_{\tilde{T}}^{-(\gamma+1)} \leq \frac{1/\delta_0}{[C_1(1+\alpha)(1+(6\Gamma)^{1/\gamma})]^\gamma} \frac{n^{\gamma-\alpha_L}}{\log^{3\gamma+1} n}, \\ h_{\tilde{T}} + 1 \leq \pi_{\tilde{T}} n / [C_1(1+\alpha) \log^4 n] \end{array} \right\}.$$

The following result establishes the typicality event (6.1) is likely to occur for any tree in \mathcal{T}_{δ_0} .

THEOREM 11. *If Assumption 2 holds then*

$$\min_{\tilde{T} \in \mathcal{T}_{\delta_0}} \min_{r \geq 1} \mathbb{P} \left(\text{Typ}_r(\tilde{T}) \right) \geq 1 - \delta_0.$$

Among the complete trees in \mathcal{T}_{δ_0} , let T^{δ_0} achieve the minimum of the following (restricted) oracle problem

$$\min_{\tilde{T} \in \mathcal{T}_{\delta_0}} \sup_{x \in A_{-\infty}^{-1}} \left(\bar{c}_{\tilde{T}(x)} + \max \left\{ \frac{c+1}{c-1} \bar{c}_{\tilde{T}(x)}, (1+2c) \|\bar{\text{cf}}(\tilde{T}(x))\|_{L,r} \right\} \right).$$

Note that the criterion above is exactly the bound in the oracle inequality developed in Theorem 5. Thus, T^{δ_0} is a context tree that yields good conditional probability estimates within trees for which we can ensure that the typicality event (6.1) is likely to hold by Theorem 11. That directly implies primitive results about the estimates (\hat{T}_n, \hat{P}_n) provided by the pruning algorithm.

THEOREM 12. *Suppose Condition AGCT and Assumption 2 hold, and let $\delta_0 \in (0, 1)$ be such that $\mathcal{T}_{\delta_0} \neq \emptyset$. Then assuming that $\text{Typ}_r(T^{\delta_0})$ and Good_m both occur, we have:*

$$\sup_{x \in A_{-\infty}^{-1}} \frac{\|d(p(\cdot|x), \hat{P}_n(\cdot|x))\|_{L,k}}{\bar{c}_{T^{\delta_0}(x)} + \max \left\{ \frac{c+1}{c-1} \bar{c}_{T^{\delta_0}(x)}, (1+2c) \|\bar{\text{cf}}(T^{\delta_0}(x))\|_{L,r} \right\}} \leq \left(1 + \frac{C_2}{\log n} \right).$$

In particular, the event above occurs with probability at least $1 - \delta - \delta_0$.

Next we provide model selection results. For trees in \mathcal{T}_{δ_0} , we show that subtrees that have well-separated paths must be selected by the estimator \hat{T}_n (Theorem 13) while trees with a small misspecification must contain the estimator \hat{T}_n (Theorem 14).

THEOREM 13. *Suppose Condition AGCT and Assumption 2 hold, and let $\tilde{T} \in \mathcal{T}_{\delta_0}$. Assume that $\tilde{T}_- \subseteq \tilde{T}$ is a subtree consisting of all nodes $w \in \tilde{T}$ such that there exist $x \in A_{-\infty}^{-1}$, $y \in A_{-\infty}^{-1}$ with $\tilde{T}(x) \succeq w$, $\tilde{T}(y) \succeq \text{par}(w)$ and:*

$$\|d(p(\cdot|x), p(\cdot|y))\|_{L,k} > \left(1 + \frac{C_2}{\log n}\right) (1+c) \left(\|\overline{\text{cf}}(\tilde{T}(x))\|_{L,r} + \|\overline{\text{cf}}(\tilde{T}(y))\|_{L,r} \right) + \bar{c}_{\tilde{T}(x)} + \bar{c}_{\tilde{T}(y)}.$$

Then if Good_m and $\text{Typ}_r(\tilde{T})$ both occur, we have that $\tilde{T}_- \subseteq \hat{T}_n$. In particular, $\mathbb{P}(\tilde{T}_- \subseteq \hat{T}_n) \geq 1 - \delta - \delta_0$.

THEOREM 14. *Suppose Condition AGCT and Assumption 2 hold. Let $\tilde{T}_+ \in \mathcal{T}_{\delta_0}$ be such that for all $x \in A_{-\infty}^{-1}$*

$$\bar{c}_{\tilde{T}_+(x)} \leq 2 \left(1 - \frac{C_2}{\log n}\right) (c-1) \|\overline{\text{cf}}(\tilde{T}_+(x))\|_{L,r}.$$

Then if Good_m and $\text{Typ}_r(\tilde{T}_+)$ both occur, we have $\hat{T}_n \subseteq \tilde{T}_+$. In particular, $\mathbb{P}(\hat{T}_n \subseteq \tilde{T}_+) \geq 1 - \delta - \delta_0$.

We briefly indicate similarities and differences between the results presented above with [15] which concerns the single-process case. The work [15] proves weak consistency in the estimation of 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 point out that their assumptions might be hard to check in practice.

Our analysis differs from theirs in several important aspects. 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 [15] 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 covers natural and interesting classes of processes (such as certain renewal processes) where geometric mixing of any kind is not available and also the use of multiple processes. Other points of the analysis are mostly incomparable due to the differences in assumptions.

6.2. Application to the parametric case. In this section we consider the behavior of our estimator in the setting where the processes $X(1), \dots, X(L)$ have a compatible context tree that is finite.

ASSUMPTION 3 (Parametric assumption). *The processes $X(1), \dots, X(L)$ are stationary and ergodic. Moreover, there exists a complete finite tree T^* and transition probabilities $p = (p_1, \dots, p_L)$ that are compatible with T^* such that:*

$$\forall 1 \leq \ell \leq L, \forall a \in A : \mathbb{P} \left(X(\ell)_0 = a \mid X(\ell)_{-\infty}^{-1} \right) = p_\ell(a \mid T^*(X(\ell)_{-\infty}^{-1})) \text{ a.s..}$$

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

$$\beta(b) = \chi e^{-\nu b} \text{ where } \chi, \nu > 0.$$

We recall that any ergodic finite-order Markov chain is exponentially ϕ -mixing, which is stronger than exponential β -mixing. Our assumption requires that the exponential β -mixing rates of the L processes be uniformly controlled.

THEOREM 15 (Rates under the Parametric Assumption). *Suppose Condition AGCT and Assumptions 1 and 3 hold simultaneously and $6\chi \leq n^{\alpha+1}$. Then there exist a constant $C_3 > 0$ depending only on $|\mathcal{S}|$ such that if:*

$$(6.3) \quad C_3 (1 + \alpha) \left(1 + \frac{\alpha + 1}{\nu} \right) \frac{h_{T^*} + 1}{\pi_{T^*}} \leq \frac{n}{\log^4 n},$$

then

$$\mathbb{P} \left(\sup_{x \in A_{-\infty}^{-1}} \frac{\|d(p(\cdot|x), \hat{P}_n(\cdot|x))\|_{L,k}}{(1+2c) \|\overline{\text{cf}}(T^*(x))\|_{L,r}} \leq \left(1 + \frac{C_2}{\log n} \right) \right) \geq 1 - \delta - \delta_0.$$

If in addition

$$d_{k,\min} \equiv \min_w \max_{w'} \left\{ \|d(p(\cdot|w), p(\cdot|w'))\|_{L,k} : w, w' \text{ leaves of } T^*, w' \succeq \text{par}(w) \right\} > 0$$

then the extra condition

$$2 \left(1 + \frac{C_2}{\log n}\right) (1+c) \max_{x \in A_{-\infty}^{-1}} \|\overline{\text{cf}}(T^*(x))\|_{L,r} < d_{k,\min},$$

implies that

$$\mathbb{P} \left(\left\{ \widehat{T}_n = T^* \right\} \cap \left\{ \sup_{x \in A_{-\infty}^{-1}} \frac{\|d(p(\cdot|x), \widehat{P}_n(\cdot|x))\|_{L,k}}{(1+2c) \|\overline{\text{cf}}(T^*(x))\|_{L,r}} \leq \left(1 + \frac{C_2}{\log n}\right) \right\} \right) \geq 1 - \delta - \delta_0.$$

To discuss the results in Theorem 15 we focus on the case that α and c are constants but $\pi_{T^*}, d_{k,\min} \ll 1$. Theorem 15 establishes that in the choice of cf for Good_∞ , we need

$$n = O \left(\left\{ \frac{1}{\pi_{T^*} d_{k,\min}^2} \log \left(\frac{1}{\pi_{T^*} d_{k,\min}^2} \right) \right\} \vee \left\{ \frac{h_{T^*}}{\pi_{T^*}} \log^4 \left(\frac{h_{T^*}}{\pi_{T^*}} \right) \right\} \right)$$

in order to correctly recover T^* with high probability. Moreover, the uniform rate of convergence of the estimation error of the conditional probabilities is bounded by $\sqrt{\log n / \pi_{T^*} n}$. By contrast, provided the number of groups L is sufficiently large, in the case cf is chosen to achieve Good_2 one needs

$$n = O \left(\left\{ \frac{1}{\pi_{T^*} d_{k,\min}^2} \log \log \left(\frac{1}{\pi_{T^*} d_{k,\min}^2} \right) \right\} \vee \left\{ \frac{h_{T^*}}{\pi_{T^*}} \log^4 \left(\frac{h_{T^*}}{\pi_{T^*}} \right) \right\} \right)$$

in order to recover T^* , and the uniform rate of convergence of the estimation error is bounded by $\sqrt{\log \log n / \pi_{T^*} n}$. This indicates the advantage of using the cf to achieve Good_2 criterion when the number L of independent processes is sufficiently large.

It is helpful to compare these results with the results on model selection for the parametric case for single-process case ($L = 1$) obtained in [8]. They considered models with compatible trees T_n^* of size increasing with the sample size n under an assumption that implies geometric φ -mixing. They show that T_n^* is correctly recovered with high probability if

$$\pi_{T_n^*}^{-1} = O \left(\frac{\sqrt{n}}{\log^{1/2+a} n} \right) \text{ and } d_{T_n^*}^{-1} = O \left(\frac{\sqrt{\pi_{T_n^*}^{-1} n}}{\log^{\frac{1+b}{2}} n} \right)$$

where $a, b > 0$ are constants and $d_{T_n^*}$ is the minimum separation between the conditional probabilities corresponding to a leaf and its parent. The latter quantity is at most the maximum separation between leaves, which is

our $d_{k,\min}$. Moreover, since they also assume that no leaves of T_n^* have null probability, so that T_n^* has at most $\pi_{T_n^*}^{-1}$ leaves. This in turn implies that T_n^* has height $h_{T_n^*} \leq \pi_{T_n^*}^{-1}$. Thus conditions in [8] imply:

$$\frac{h_{T_n^*}}{\pi_{T_n^*}} = O\left(\frac{n}{\log^{1+2a} n}\right) \text{ and } \frac{1}{\pi_{T_n^*} d_{k,\min}^2} = O\left(\frac{n}{\log^{1+b} n}\right).$$

Those conditions match our own conditions up to the logarithmic terms. Thus the result presented here nearly implies the perfect model selection result in [8], even with our pessimistic estimates on $h_{T_n^*}$ and $d_{k,\min}$, and offers improvements in situations where $h_{T_n^*}/\pi_{T_n^*} \ll 1/\log^3 n$ and/or $d_{k,\min} \ll d_{T_n^*}$. Additionally, we also obtain improvements in the group setting which was not considered in the context tree literature before.

6.3. Chains of infinite order with complete connections. We now consider processes which do *not* satisfy the parametric assumption, but which can be well-approximated by order- k Markov chains. We focus on a particularly well-understood case where $X(1), \dots, X(L)$ are known to be ϕ -mixing [10].

We first need some definitions. Let:

$$q_{\min} \equiv \min_{1 \leq \ell \leq L, a \in A, x \in A_{-\infty}^{-1}} p_{\ell}(a|x).$$

Also define the *continuity rates*:

$$\lambda_{\ell}(b) \equiv 1 - \min_{w \in A_{-b}^{-1}} \sum_{a \in A} \inf_{x \in A_{-\infty}^{-1} : x_{-b}^{-1} = w} p_{\ell}(a|x) \quad (b \in \mathbb{N} \setminus \{0\}, 1 \leq \ell \leq L).$$

ASSUMPTION 4 (Assumption of chains with complete connections). *Assumption 1 holds, and assume $q_{\min} > 0$ and for some $\gamma > 2\alpha_L + 1$*

$$\lambda_{\ell}(b) \leq \Gamma_0 b^{-1-\gamma} \quad (b \in \mathbb{N}).$$

To see how the assumption relates to group context tree models, consider the “canonical approximation” of $X(\ell)$ by a order- b Markov chain, with transition probabilities:

$$\tilde{p}_{\ell}(a|T_b(x)) = \mathbb{P}\left(X(\ell)_0 = a \mid X_{-b}^{-1} = w\right)$$

where T_b denotes the tree that contains all suffixes of length at most b .

One can check that the distance $d_{\mathcal{S}}$ between $\tilde{p}_{\ell}(\cdot|T_b(x))$ and $p_{\ell}(a|x)$ is at most $\lambda_{\ell}(b)$. It follows that, the faster $\lambda_{\ell}(b)$ decays, the better $X(\ell)$ can be approximated by Markov chains of finite order.

THEOREM 16 (Rates for Chains with Infinite Connections). *Suppose Condition AGCT and Assumption 4 hold. There exist constants $C_4, C_5 > 0$, depending only on $c, |\mathcal{S}|, q_{\min}, \Gamma_0$ and γ such that if*

$$(6.4) \quad \frac{n}{\log^{18} n} \geq C_4 (1 + \alpha)^6,$$

then

$$\mathbb{P} \left(\sup_{x \in A_{-\infty}^{-1}} \left\| d(\widehat{P}_n(\cdot|x), p(\cdot|x)) \right\|_{L,k} \leq \frac{C_5}{\log^{\gamma+1} n} \right) \geq 1 - \delta - n^{-\frac{\gamma-1-2\alpha_L}{2}}.$$

For example, this probability is $1 - O(n^{-2})$ if $\alpha_L \leq 1$ (i.e., $L \leq n$), $\delta = n^{-2}$ and $\gamma \geq 7$. In contrast to the previous section, the difference between the ℓ_1 and ℓ_2 cases is not apparent in the simplified bound presented above. We also note that, with high probability, the estimated tree only contains strings of size $O(\log n)$: $q_{\min} > 0$ implies that any string w has stationary probability $\leq (1 - q_{\min})^{|w|}$, so no strings of length much larger than $\log n$ will ever be seen in the sample. No statement about “lower bounds” for \widehat{T}_n in the spirit of Theorem 13 can be made at this level of generality.

7. Simulations. In this section we conduct Monte Carlo experiments to assess the finite-sample performance of the proposed estimator. We use two different designs for the true context tree T^* in these experiments: (i) a full binary Markov chain of order 3, and (ii) a sparse binary VLMC with infinite length associated with a renewal process. The associated context trees are displayed in Table 1. The former model corresponds to a parametric model with well separated conditional probabilities. The latter corresponds to an infinite context tree induced by a renewal process. These designs are two extreme cases to help illustrate the performance of the estimator on balanced and unbalanced context trees. For each design we consider the size of the group to be $L = 1, 10, 100$, various sample sizes n , and two different choices of the regularization parameter, $c = 1.01, 0.5$. In all simulations we used $d_\ell = \|\cdot\|_\infty$, $k = 1$, $r = 2$, $m = 2$, and set the confidence level with $1 - \delta = 0.95$.

Table 2 displays the model selection performance of the proposed algorithm for the full binary Markov chain of order 3 when the parameter c is set to 1.01 and 0.5. In the case of $c = 1.01$ that follows the theoretical recommendation of the previous section, in every instance the estimated tree \widehat{T}_n was contained in the true context tree T^* confirming our theoretical results. Moreover, the estimated context tree contained a full binary Markov chain

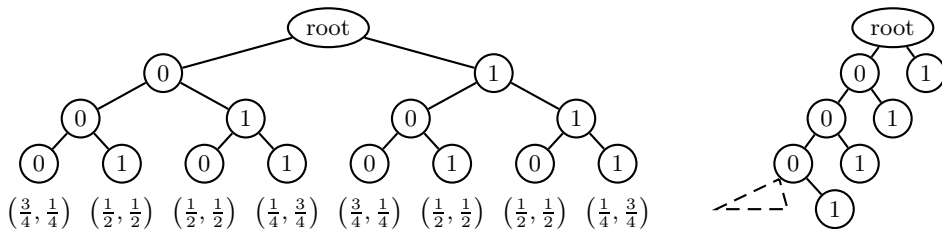


TABLE 1

The context trees above illustrate the two models used in our simulations. The left context tree correspond to a full binary Markov chain of order 3. The right context tree correspond a process with infinite memory associated with a renewal process.

of order 2 in most instances. In the larger sample size with 100 groups, we achieved perfect recovery of the model. When we set $c = 0.5$ additional nodes not in T^* are occasionally included (the average number of extra nodes is displayed in the last row of the table labeled as “extra”). If multiple groups are used in the estimation, the number of extra nodes selected was smaller.

The renewal process is defined by the independent times t_i 's between observing two 1's. We specified the random variable t_i such as $P(t_i = k) = 1/[(2 \log 2 - 1)k(4k^2 - 1)]$. Stationarity requires that the first time be drawn from a different distribution, see [13, 17] for details. We note that the polynomial decay of the tail suggests potentially long estimated trees. Table 3 displays the model selection performance of the proposed algorithm for the renewal process when the parameter c is set to 1.01 and 0.5. In the case of $c = 1.01$ that follows the theoretical recommendation of the previous section, in every instance the estimated tree \hat{T}_n was contained in the true context tree T^* confirming our theoretical results. As expected, as the sample size increases the estimated context tree also increases chasing the infinite true context tree. When we set $c = 0.5$ additional nodes not in T^* are occasionally included (the average number of extra nodes is displayed in the last row of the table labeled as “extra”). Nonetheless, when multiple groups are used in the estimation, no node outside of the true context tree are selected.

8. Linguistic rhythm differences between European and Brazilian Portuguese. In this section we revisit the application and the data considered in [16] regarding the linguistic features underlying the European Portuguese and Brazilian Portuguese languages. 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

Node	Probability of selection with parameter $c = 1.01$								
	n=1000			n=2500			n=5000		
	L=1	L=10	L= 100	L=1	L=10	L= 100	L=1	L=10	
000	0.0	0.0	0.0	0.02	0.48	1.0	0.68	1.0	
100	0.0	0.0	0.0	0.02	0.48	1.0	0.68	1.0	
010	0.0	0.0	0.0	0.03	0.17	0.83	0.57	1.0	
110	0.0	0.0	0.0	0.03	0.17	0.83	0.57	1.0	
001	0.0	0.0	0.0	0.03	0.11	0.09	0.53	1.0	
101	0.0	0.0	0.0	0.03	0.11	0.09	0.53	1.0	
011	0.0	0.0	0.0	0.04	0.42	1.0	0.69	1.0	
111	0.0	0.0	0.0	0.04	0.42	1.0	0.69	1.0	
00	0.36	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
10	0.36	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
01	0.4	0.98	1.0	1.0	1.0	1.0	1.0	1.0	
11	0.4	0.98	1.0	1.0	1.0	1.0	1.0	1.0	
0	0.66	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
1	0.66	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
root	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
extra	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Node	Probability of selection with parameter $c = 0.5$								
	n=1000			n=2500			n=5000		
	L=1	L=10	L=100	L=1	L=10	L=100	L=1	L=10	
000	0.92	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
100	0.92	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
010	0.83	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
110	0.83	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
001	0.84	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
101	0.84	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
011	0.91	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
111	0.91	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
00	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
01	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
11	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
root	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
extra	4.22	0.0	0.0	8.67	0.25	0.0	27.55	11.02	

TABLE 2

The table illustrates the model selection performance for selecting nodes of the true context tree in the full binary Markov chain of order 3.

four rhythmic features and an additional feature representing the end of an article. The four rhythmic features represent: non-stressed, non prosodic word initial syllable (0); stressed, non prosodic word initial syllable (1); non-stressed, prosodic word initial syllable (2); and stressed prosodic word initial syllable (3).

Node	Probability of selection with parameter $c = 1.01$								
	n=5000			n=10000			n=50000		
	L=1	L=10	L=100	L=1	L=10	L=100	L=1	L=10	
others	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
00000000	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	
10000000	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	
0000000	0.0	0.0	0.0	0.0	0.0	0.0	0.03	0.0	
1000000	0.0	0.0	0.0	0.0	0.0	0.0	0.03	0.0	
000000	0.0	0.0	0.0	0.0	0.0	0.0	0.46	0.31	
100000	0.0	0.0	0.0	0.0	0.0	0.0	0.46	0.31	
00000	0.0	0.0	0.0	0.01	0.0	0.0	0.97	1.0	
10000	0.0	0.0	0.0	0.01	0.0	0.0	0.97	1.0	
0000	0.02	0.0	0.75	0.37	0.92	1.0	1.0	1.0	
1000	0.02	0.0	0.75	0.37	0.92	1.0	1.0	1.0	
000	0.72	1.0	1.0	0.99	1.0	1.0	1.0	1.0	
100	0.72	1.0	1.0	0.99	1.0	1.0	1.0	1.0	
00	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
root	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
extra	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Node	Probability of selection with parameter $c = 0.5$								
	n=5000			n=10000			n=50000		
	L=1	L=10	L=100	L=1	L=10	L=100	L=1	L=10	
others	0.18	0.0	0.0	0.02	0.0	0.0	1.84	2.76	
00000000	0.03	0.0	0.0	0.03	0.0	0.0	0.77	1.0	
10000000	0.03	0.0	0.0	0.03	0.0	0.0	0.77	1.0	
0000000	0.04	0.0	0.0	0.06	0.02	0.0	0.99	1.0	
1000000	0.04	0.0	0.0	0.06	0.02	0.0	0.99	1.0	
000000	0.08	0.0	0.0	0.36	0.53	0.72	1.0	1.0	
100000	0.08	0.0	0.0	0.36	0.53	0.72	1.0	1.0	
00000	0.28	0.42	0.23	0.73	1.0	1.0	1.0	1.0	
10000	0.28	0.42	0.23	0.73	1.0	1.0	1.0	1.0	
0000	0.78	1.0	1.0	0.98	1.0	1.0	1.0	1.0	
1000	0.78	1.0	1.0	0.98	1.0	1.0	1.0	1.0	
000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
100	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
00	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
root	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
extra	4.36	0.0	0.0	6.5	0.0	0.0	3.54	0.0	

TABLE 3

The table illustrates the model selection performance for selecting nodes of the true context tree in the binary Markov chain induced by renewal process X_t .

In [16], for each newspaper, the 40 days sample is concatenated into a single string containing respectively a sequence of 105326 and 97750 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.

We propose to account for the possible heterogeneity on the conditional probabilities and consider each year as a group in the group context tree model. Thus we allow for year specific conditional probabilities. Figure 2 displays the estimated context trees. As in [16], we find that the European Portuguese has a more complex context tree possibly reflecting the changes the language suffered during the 18th century. Both context trees are similar to the trees found in [16] confirming their finding even in the presence of heterogeneity.

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 the 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 oracle inequalities for model selection and estimation of conditional probabilities. Two methodological applications, discrete dynamic stochastic programming and discrete choice models, motivated the proposal of the group context tree model. In these applications we are interested 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. We also propose and analyze data-driven choices of the penalty choices for the regularization, and study its typical behavior under β -mixing conditions.

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 accounting for possible heterogeneity in the sample. Our results fully support previous findings of the literature.

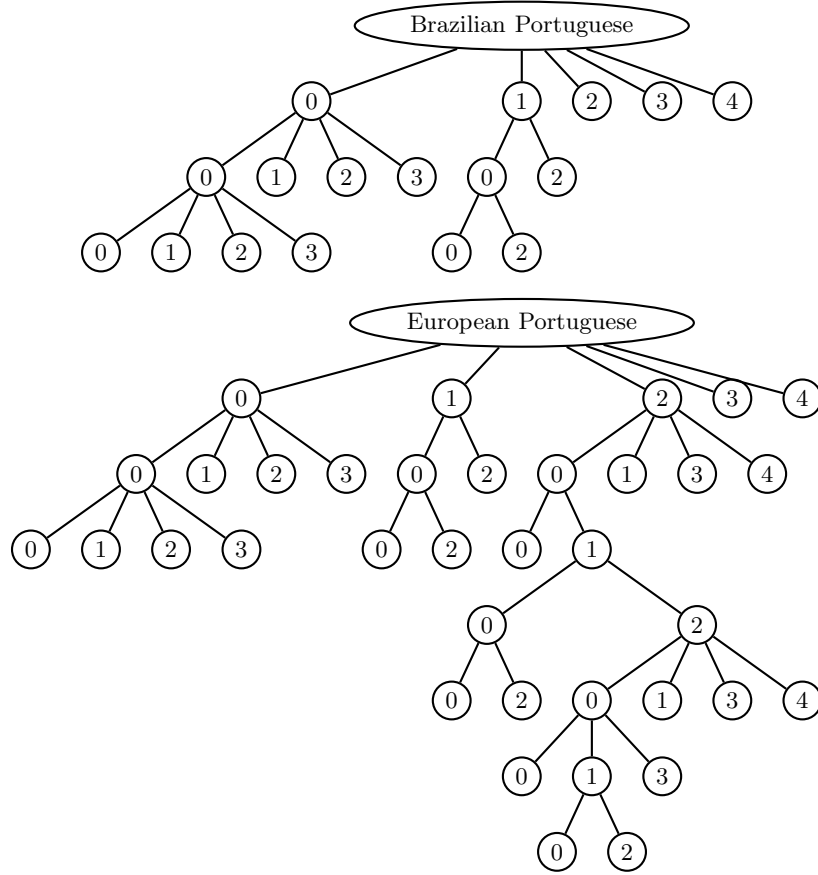


FIG 2. Estimated context trees for the Brazilian Portuguese and European Portuguese languages based accounting for heterogeneity in different years.

APPENDIX A: TECHNICAL PROPERTIES

We start by establishing technical results that follows from the definition of the pruning algorithm. We begin with an alternative characterization of \hat{T}_n .

PROPOSITION 1. *The estimated tree \hat{T}_n equals the smallest tree contained in E_n which contains all $w \in E_n$ with $\text{CanRmv}(w) = 0$.*

PROOF OF PROPOSITION 1. Let S_n denote the aforementioned “smallest tree”. Clearly, $S_n \subset \hat{T}_n$, as any w with $\text{CanRmv}(w) = 0$ will not be removed from \hat{T}_n in PruneTree. To prove that $E_n \setminus S_n \subset E_n \setminus \hat{T}_n$, we will use the following claim:

CLAIM 1. *If $v \in E_n \setminus S_n$, then $\forall w \in E_n : w \succeq v \Rightarrow \text{CanRmv}(w) = 1$.*

PROOF OF CLAIM. In contrapositive form, if some $w \succeq v$ satisfies that $\text{CanRmv}(w) = 0$, that w belongs to S_n by definition, and then $v \in S_n$ because S_n is a tree and $v \preceq w$. \square

One may use induction starting from the leaves to deduce that, if $v \in E_n$ is such that the conclusion of the Claim holds, it will be removed from \widehat{T}_n at some stage of PruneTree. We deduce $E_n \setminus S_n \subset E_n \setminus \widehat{T}_n$. \square

It turns out that, except for the root, any node of the estimated tree \widehat{T}_n must be the closely connected with two nodes that yields substantially different probability distributions.

PROPOSITION 2. *Suppose $v \in \widehat{T}_n \setminus \{e\}$. Then there exist $w', w'' \in E_n$ with $w' \succeq \text{par}(v)$, $w'' \succeq v$ and*

$$c \left[\|\text{cf}(w')\|_{L,r} + \|\text{cf}(w'')\|_{L,r} \right] < \|d(\hat{p}_n(\cdot|w'), \hat{p}_n(\cdot|w''))\|_{L,k}.$$

PROOF OF PROPOSITION 2. Assume (to get a contradiction) that no such w', w'' exist. In that case one can easily check that $\text{CanRmv}(w) = 1$ for all $w \succeq v$. In particular, the subtree of E_n obtained by removing v and all of its descendants contains all u with $\text{CanRmv}(u) = 0$. Proposition 1 then implies $v \notin \widehat{T}_n$, which contradicts the assumptions of the present Proposition and finishes the proof. \square

Finally, the following result formally states the compatibility between the tree structure \widehat{T}_n and the probability distributions $\hat{P}_n(\cdot|x)$ which follows immediately from the pruning definition.

PROPOSITION 3. *Let $x, y \in A_{-\infty}^{-1}$ satisfy $\widehat{T}_n(x) = \widehat{T}_n(y)$. Then $\hat{P}_n(\cdot|x) = \hat{P}_n(\cdot|y)$.*

APPENDIX B: PROOFS OF SECTION 3

PROOF OF THEOREM 1. We prove the case that r, m , and k are finite (the other case follows similarly). If $\min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) = 0$ by definition we have $\bar{p}_{n,\ell}(a|w) = \hat{p}_{n,\ell}(a|w) = 1/|A|$ and the result follows. Otherwise, using the definitions of $\|\cdot\|_{L,k}$ and $\|\cdot\|_{L,r}$, and Holder's inequality (since $m > k$) we

have

$$\begin{aligned}
\|d(\hat{p}_n(\cdot|w), \bar{p}_n(\cdot|w))\|_{L,k}^k &= \frac{1}{L} \sum_{\ell=1}^L d_\ell^k(\hat{p}_{n,\ell}(\cdot|w), \bar{p}_{n,\ell}(\cdot|w)) \\
&= \frac{1}{L} \sum_{\ell=1}^L \frac{d_\ell^k(\hat{p}_{n,\ell}(\cdot|w), \bar{p}_{n,\ell}(\cdot|w))}{\text{cf}_\ell^k(w)} \text{cf}_\ell^k(w) \\
&\leq \left(\frac{1}{L} \sum_{\ell=1}^L \frac{d_\ell^m(\hat{p}_{n,\ell}(\cdot|w), \bar{p}_{n,\ell}(\cdot|w))}{\text{cf}_\ell^m(w)} \right)^{\frac{k}{m}} \left(\frac{1}{L} \sum_{\ell=1}^L \text{cf}_\ell^{\frac{km}{m-k}}(w) \right)^{\frac{m-k}{m}} \\
&= \left\| \left\{ \frac{d_\ell(\bar{p}_{n,\ell}(\cdot|w), \hat{p}_{n,\ell}(\cdot|w))}{\text{cf}_\ell(w)} \right\}_{\ell=1}^L \right\|_{L,m}^k \|\text{cf}(w)\|_{L, \frac{km}{m-k}}^k.
\end{aligned}$$

Thus, if Good_m occurs, for any $r \geq km/(m-k)$ we have

$$\|d(\hat{p}_n(\cdot|w), \bar{p}_n(\cdot|w))\|_{L,k} \leq \|\text{cf}(w)\|_{L,r}$$

since $\|v\|_{L,r'} \leq \|v\|_{L,r}$ if $r' \leq r$. \square

PROOF OF THEOREM 2. We need to show that no node outside of T^* belongs to \hat{T}_n . A string $z \in A^*$ lies outside of T^* if there exists a leaf w of T^* such that $w \prec z$, $w \neq z$. Notice that, if such a w exists, we have $\bar{p}_n(\cdot|w') = \bar{p}_n(\cdot|w'')$ for all $w', w'' \succeq w$. Thus, if Good_m holds, by Theorem 1 we may deduce that:

$$\begin{aligned}
\|d(\hat{p}_n(\cdot|w'), \hat{p}_n(\cdot|w''))\|_{L,k} &\leq \|d(\hat{p}_n(\cdot|w'), \bar{p}_n(\cdot|w'))\|_{L,k} + \|d(\bar{p}_n(\cdot|w''), \hat{p}_n(\cdot|w''))\|_{L,k} \\
&\leq \|\text{cf}(w')\|_{L,r} + \|\text{cf}(w'')\|_{L,r}.
\end{aligned}$$

This implies that the entire subtree of E_n rooted at w is pruned in the criterion (2.4) since $c > 1$, and therefore $z \notin \hat{T}_n$. \square

PROOF OF THEOREM 3. Let $x \in A_{-\infty}^{-1}$ and set $w = T(x)$ and $\hat{w} = \hat{T}_n(x)$. Since $\|\text{cf}(w)\|_{L,r}$ is monotone in w , if $|\hat{w}| \leq |w|$ we have $\|\text{cf}(\hat{w})\|_{L,r} \leq \|\text{cf}(w)\|_{L,r}$.

Next assume that $|\hat{w}| > |w|$ and let $\tilde{w} = \text{par}(\hat{w})$. By Proposition 2 there exist $w', w'' \in E_n$ with $w' \succeq \tilde{w} \succeq w$, $w'' \succeq \hat{w} \succeq w$ and

$$c \left[\|\text{cf}(w')\|_{L,r} + \|\text{cf}(w'')\|_{L,r} \right] < \|d(\hat{p}_n(\cdot|w'), \hat{p}_n(\cdot|w''))\|_{L,k}.$$

On the other hand, we *claim* that

CLAIM 2. *In Good_m , for all $w', w'' \in A^*$ with $w', w'' \succeq w$,*

$$\|d(\hat{p}_n(\cdot|w'), \hat{p}_n(\cdot|w''))\|_{L,k} \leq 2c_w + \|\text{cf}(w')\|_{L,r} + \|\text{cf}(w'')\|_{L,r}.$$

The Claim implies:

$$c \left[\|\mathbf{cf}(w')\|_{L,r} + \|\mathbf{cf}(w'')\|_{L,r} \right] < 2c_w + \|\mathbf{cf}(w')\|_{L,r} + \|\mathbf{cf}(w'')\|_{L,r},$$

which (since $c > 1$) is the same as:

$$\|\mathbf{cf}(w')\|_{L,r} + \|\mathbf{cf}(w'')\|_{L,r} < \frac{2c_w}{c-1}.$$

It follows that

$$\|\mathbf{cf}(\widehat{w})\|_{L,r} \leq \frac{2c_w}{c-1} - \|\mathbf{cf}(w)\|_{L,r},$$

since $\mathbf{cf}_\ell(\cdot)$ is increasing, $w \preceq w'$ and $\widehat{w} \preceq w''$. Thus we only need to prove the Claim in order to finish the proof of the present Theorem.

To prove the Claim, fix some $w' \in E_n$ with $w' \succeq w$ and write $s \equiv |w'|$. We observe that:

$$\begin{aligned} \text{(B.1)} \quad \left\| d(\bar{p}_n(\cdot|w), \bar{p}_n(\cdot|w')) \right\|_{L,k} &\leq \left\| \left\{ \frac{\sum_{i=s+1}^n \chi_{\{X_{i-s}^{i-1}(\ell)=w'\}} d_\ell(\bar{p}_n, \ell(\cdot|w), p_\ell(\cdot|X_{-\infty}^{i-1}(\ell)))}{N_{n-1, \ell}(w')} \right\}_{\ell=1}^L \right\|_{L,k} \\ &\leq \sup_{\{z(\ell) : z_{-s}^{-1}(\ell)=w\}_{\ell=1}^L} \left\| \left\{ d_\ell(\bar{p}_n, \ell(\cdot|w), p_\ell(\cdot|z(\ell))) \right\}_{\ell=1}^L \right\|_{L,k} \\ \text{(} w = T(x) \text{ and Rem. 1)} &\leq \sup_{\{z(\ell) : T(z(\ell))=w\}_{\ell=1}^L} \left\| \left\{ d_\ell(\bar{p}_n, \ell(\cdot|w), p_\ell(\cdot|z(\ell))) \right\}_{\ell=1}^L \right\|_{L,k} \\ &\leq c_w = c_{T(x)}. \end{aligned}$$

Recall from Theorem 1 that if Good_m holds, $\|d(\hat{p}_n(\cdot|w'), \bar{p}_n(\cdot|w'))\|_{L,k} \leq \|\mathbf{cf}(w')\|_{L,r}$ for all $w' \in A^*$, hence by the triangle inequality:

$$\text{(B.2)} \quad \begin{aligned} &\text{In } \text{Good}_m, \forall w' \in A^* : \\ &w' \succeq w \Rightarrow \|d(\hat{p}_n(\cdot|w'), \hat{p}_n(\cdot|w))\|_{L,k} \leq c_{T(x)} + \|\mathbf{cf}(w')\|_{L,r}. \end{aligned}$$

Comparing $w', w'' \succeq w$ via the triangle inequality and (B.2) finishes the proof of the Claim. \square

PROOF OF THEOREM 4. First apply Theorem 3 with $|\widehat{T}_n(x)| > |T(x)|$ so that

$$\left\| \mathbf{cf}(\widehat{T}_n(x)) \right\|_{L,r} \leq \frac{2c_{T(x)}}{c-1}.$$

Next, by condition RL, we have

$$\left\| \mathbf{cf}(\widehat{T}_n(x)) \right\|_{L,r} \geq \|\mathbf{cf}(T(x))\|_{L,r} \left(\frac{1}{\kappa} \right)^{\frac{|\widehat{T}_n(x)| - |T(x)|}{k}}$$

since $\mathbf{cf}_\ell(\widehat{T}_n(x)) \leq 1$ so that $\left\| \mathbf{cf}(\widehat{T}_n(x)) \right\|_{L,r} \leq 1$. The result follows by combining both inequalities. \square

PROOF OF THEOREM 5. By the triangle inequality and the definition of $c_{T(x)}$

$$\left\| d(\hat{p}_n(\cdot|\hat{T}_n(x)), p(\cdot|x)) \right\|_{L,k} \leq \left\| d(\hat{p}_n(\cdot|\hat{T}_n(x)), \bar{p}_n(\cdot|T(x))) \right\|_{L,k} + c_{T(x)}$$

so that the second relation follows from the first.

To prove the first relation, note that under Good_m , by Theorem 1 we have

$$(B.3) \quad \left\| d(\bar{p}_n(\cdot|T(x)), \hat{p}_n(\cdot|T(x))) \right\|_{L,k} \leq \|\text{cf}(T(x))\|_{L,r}.$$

We divide the rest of the analysis into three cases.

Case 0: $|\hat{T}_n(x)| = |T(x)|$. The result follows from the relation above since $\hat{T}_n(x) = T(x)$.

Case 1: $|\hat{T}_n(x)| < |T(x)|$. In this case by the triangle inequality and (B.3) we have

$$\begin{aligned} \left\| d(\bar{p}_n(\cdot|T(x)), \hat{p}_n(\cdot|\hat{T}_n(x))) \right\|_{L,k} &\leq \left\| d(\hat{p}_n(\cdot|\hat{T}_n(x)), \hat{p}_n(\cdot|T(x))) \right\|_{L,k} + \left\| d(\hat{p}_n(\cdot|T(x)), \bar{p}_n(\cdot|T(x))) \right\|_{L,k} \\ &\leq \left\| d(\hat{p}_n(\cdot|\hat{T}_n(x)), \hat{p}_n(\cdot|T(x))) \right\|_{L,k} + \|\text{cf}(T(x))\|_{L,r}. \end{aligned}$$

Next note that $T(x)$ was pruned since $|\hat{T}_n(x)| < |T(x)|$. Therefore,

$$\left\| d(\hat{p}_n(\cdot|\hat{T}_n(x)), \hat{p}_n(\cdot|T(x))) \right\|_{L,k} \leq c \left[\left\| \text{cf}(\hat{T}_n(x)) \right\|_{L,r} + \|\text{cf}(T(x))\|_{L,r} \right].$$

Thus we have

$$\begin{aligned} \left\| d(\bar{p}_n(\cdot|T(x)), \hat{p}_n(\cdot|\hat{T}_n(x))) \right\|_{L,k} &\leq c \left\| \text{cf}(\hat{T}_n(x)) \right\|_{L,r} + (1+c) \|\text{cf}(T(x))\|_{L,r} \\ &\leq (1+2c) \|\text{cf}(T(x))\|_{L,r} \end{aligned}$$

since $\left\| \text{cf}(\hat{T}_n(x)) \right\|_{L,r} \leq \|\text{cf}(T(x))\|_{L,r}$ because $|\hat{T}_n(x)| < |T(x)|$.

Case 2: $|T(x)| < |\hat{T}_n(x)|$. In this case, notice that by (B.1) we have

$$\left\| d(\bar{p}_n(\cdot|T(x)), \bar{p}_n(\cdot|\hat{T}_n(x))) \right\|_{L,k} \leq c_{T(x)}.$$

Hence, under Good_m , using the triangle inequality and Theorems 1 and 3, we have

$$\begin{aligned} \left\| d(\hat{p}_n(\cdot|\hat{T}_n(x)), \bar{p}_n(\cdot|T(x))) \right\|_{L,k} &\leq \left\| d(\hat{p}_n(\cdot|\hat{T}_n(x)), \bar{p}_n(\cdot|\hat{T}_n(x))) \right\|_{L,k} + \left\| d(\bar{p}_n(\cdot|\hat{T}_n(x)), \bar{p}_n(\cdot|T(x))) \right\|_{L,k} \\ &\leq \left\| \text{cf}(\hat{T}_n(x)) \right\|_{L,r} + c_{T(x)} \\ &\leq \frac{2c_{T(x)}}{c-1} + c_{T(x)} = \frac{c+1}{c-1} c_{T(x)}. \end{aligned}$$

□

APPENDIX C: PROOFS OF SECTION 4

Consider a collection of subsets $\mathcal{S} \subseteq 2^A$. We consider the pseudo-metric on the simplex Δ^A given by:

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

For each process $\ell = 1, \dots, L$ and $S \in \mathcal{S}$, we have empirical transition probabilities:

$$\hat{p}_{n,\ell}(S|w) = \frac{N_{n,\ell}(wS)}{N_{n-1,\ell}(w)} \text{ where } N_{n,\ell}(wS) = \sum_{a \in S} N_{n,\ell}(wa)$$

and the ‘‘oracle probabilities’’:

$$\bar{p}_{n,\ell}(S|w) = \frac{\sum_{i=1}^n \chi_{\{X_{i-|w|}^i(\ell)=w\}} p_\ell(S|X_{-\infty}^i(\ell))}{N_{n-1,\ell}(w)}$$

if $\min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) > 0$. Both probabilities are defined as $|S|/|A|$ when $\min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) = 0$. We will study these quantities using the martingale framework from Appendix E. A simple calculation (omitted) proves the following proposition.

PROPOSITION 4. *Given $w \in A^*$, $S \in \mathcal{S}$ and $1 \leq \ell \leq L$, write $X_{0,\ell}(wS) = 0$ and for $1 \leq t \leq n$:*

$$X_{t,\ell}(wS) \equiv N_{t,\ell}(wS) - \sum_{i=1}^t \chi_{\{X_{i-|w|}^i(\ell)=w\}} p_\ell(S|X_{-\infty}^i(\ell)).$$

Then this is a martingale with filtration $\sigma(X_{-\infty}^t(\ell))_{t \geq 0}$ and quadratic variation process

$$\begin{aligned} V_{t,\ell} &\equiv V_{t,\ell}(w) = \sum_{j=1}^t E[|X_{j,\ell} - X_{j-1,\ell}|^2 | \sigma(X_{-\infty}^t(\ell))] \\ &\leq N_{t-1,\ell}(w) \bar{p}_{n,\ell}(S|w) (1 - \bar{p}_{n,\ell}(S|w)). \end{aligned}$$

Proposition 4 will be useful in our calculations because $X_{t,\ell}$ defined above satisfies:

$$(C.1) \quad d_\ell(\hat{p}_{n,\ell}(\cdot|w), \bar{p}_{n,\ell}(\cdot|w)) = \sup_{S \in \mathcal{S}} \frac{|X_{n,\ell}(wS)|}{N_{n-1,\ell}(w)}.$$

Note that the proof of Theorems 6 and 7 follows directly from Theorem 8. The proof relies on martingale inequalities developed in the Appendix E.

For $\gamma > 1$ define

$$(C.2) \quad c_\ell(w) = \gamma \sqrt{\frac{2\bar{\sigma}_\ell^2(w)}{N_{n-1,\ell}(w)}} \sqrt{\log \left\{ \frac{4|\mathcal{S}|(\log n) \log(n^2 L / \delta)}{\log \log n} \right\} + 2 \log \{2 + \log_\gamma[\bar{\sigma}_\ell^2(w) N_{n-1,\ell}(w)]\}},$$

if

$$\bar{\sigma}_\ell^2(w)N_{n-1,\ell}(w) \geq \gamma^{i_0} \geq \frac{2 \log(n^2|\mathcal{S}|/\delta) + 2 \log[(1+i_0)(2+i_0)]}{\log^2(2-(1/\gamma))},$$

where i_0 be the smallest such integer, and
(C.3)

$$c_\ell(w) = \sqrt{\frac{2\gamma}{N_{n-1,\ell}(w)}} \sqrt{\log \left\{ \frac{4|\mathcal{S}|(\log n) \log(n^2L/\delta)}{\log \log n} \right\} + 2 \log \{2 + \log_\gamma[N_{n-1,\ell}(w)]\}},$$

otherwise. The results stated in Section 4 correspond to choosing $\gamma = 2$. It follows that $\text{cf}_\ell^{\bar{\sigma}}(w) \geq c_\ell(w)$ with $\gamma = 2$.

PROOF OF THEOREM 8, Good $_\infty$. For notational convenience let $d_\ell(w) = d_\ell(\hat{p}_{n,\ell}(\cdot|w), \bar{p}_{n,\ell}(\cdot|w))$, and the event $E_w = \{\min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) > 0\}$.

$$\mathbb{P} \left(\exists w \in A^* : \max_{\ell=1,\dots,L} \frac{d_\ell(w)}{\text{cf}_\ell(w)} > 1 \right) \leq \sum_{w \in A^*} \mathbb{P} \left(\max_{\ell=1,\dots,L} \frac{d_\ell(w)}{\text{cf}_\ell(w)} > 1 | E_w \right) \mathbb{P}(E_w).$$

Using (C.1), the union bound and Lemma 5 applied (twice) to each $S \in \mathcal{S}$, and for every $\ell = 1, \dots, L$, we have that

$$\mathbb{P} \left(\max_{\ell=1,\dots,L} \frac{d_\ell(w)}{\text{cf}_\ell(w)} > 1 | E_w \right) \leq \sum_{S \in \mathcal{S}} \sum_{\ell=1}^L \mathbb{P} \left(\frac{|X_{n,\ell}(wS)|}{N_{n-1,\ell}(w)\text{cf}_\ell(w)} > 1 | E_w \right) \leq \frac{2\delta}{n^2}.$$

The result follows by noting that $\sum_{w \in A^*} \mathbb{P}(E_w) \leq n^2/2$ where the last expression is the maximum number of different substrings of a string of length n (note that E_w requires the substring to appear in all L strings). \square

Before proceeding to the second result of the section we need a technical lemma.

LEMMA 2. *Assume that $\text{cf}(w)$ is a random variable such that for some $M \geq 1$ we have that for every $\ell = 1, \dots, L$*

$$E \left[\frac{d_\ell^2(\hat{p}_{n,\ell}(\cdot|w), \bar{p}_{n,\ell}(\cdot|w))}{\text{cf}_\ell^2(w)} \mid d_\ell(\hat{p}_{n,\ell}(\cdot|w), \bar{p}_{n,\ell}(\cdot|w)) \leq \sqrt{M}\text{cf}_\ell(w), N_{n-1,\ell}(w) > 0 \right] \leq 1 \text{ and}$$

$$\mathbb{P} \left(d_\ell(\hat{p}_{n,\ell}(\cdot|w), \bar{p}_{n,\ell}(\cdot|w)) > \sqrt{M}\text{cf}_\ell(w) | N_{n-1,\ell}(w) > 0 \right) \leq \delta^M.$$

Then

$$\begin{aligned} & \mathbb{P} \left(\exists w \in A^* : \min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) > 0, \frac{1}{L} \sum_{\ell=1}^L \frac{d_\ell^2(\hat{p}_{n,\ell}(\cdot|w), \bar{p}_{n,\ell}(\cdot|w))}{\text{cf}_\ell^2(w)} > 1 + h \right) \\ & \leq \frac{n^2}{2} \left(\exp \left(-\frac{1}{2} \frac{h^2 L}{M(1+h/3)} \right) + L\delta^M \right). \end{aligned}$$

PROOF. For notational convenience let $d_\ell(w) = d_\ell(\hat{p}_{n,\ell}(\cdot|w), \bar{p}_{n,\ell}(\cdot|w))$ and the events $E_{\ell,w} = \{N_{n-1,\ell}(w) > 0\}$ and $E_w = \{\min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) > 0\}$. First note that

$$\mathbb{P}\left(\exists w \in A^* : E_w, \frac{1}{L} \sum_{\ell=1}^L \frac{d_\ell^2(w)}{cf_\ell^2(w)} > 1 + h\right) \leq \sum_{w \in A^*} \mathbb{P}\left(\frac{1}{L} \sum_{\ell=1}^L \frac{d_\ell^2(w)}{cf_\ell^2(w)} > 1 + h | E_w\right) \cdot \mathbb{P}(E_w).$$

Then we have that for each $w \in A^*$ and $M \geq 0$

$$\begin{aligned} \mathbb{P}\left(\frac{1}{L} \sum_{\ell=1}^L \frac{d_\ell^2(w)}{cf_\ell^2(w)} > 1 + h | E_w\right) &\leq \mathbb{P}\left(\frac{1}{L} \sum_{\ell=1}^L \frac{d_\ell^2(w)}{cf_\ell^2(w)} > 1 + h | E_w, d_\ell^2(w) \leq M cf_\ell^2(w)\right) + \\ &\quad + \mathbb{L} \max_{1 \leq \ell \leq L} \mathbb{P}(d_\ell(w) > \sqrt{M} cf_\ell(w) | E_{\ell,w}). \end{aligned}$$

Note that condition on the events $E_{\ell,w}$ and $\{d_\ell^2(w) \leq M cf_\ell^2(w)\}$, we have that the variable $Z_\ell := (d_\ell^2(w)/cf_\ell^2(w)) - 1$ is such that $E[Z_\ell] \leq 0$, $|Z_\ell| \leq M$, and $E[Z_\ell^2] \leq M$. Then by Bernstein's inequality we have

$$\mathbb{P}\left(\frac{1}{L} \sum_{\ell=1}^L Z_\ell > h \mid E_w, d_\ell^2(w) \leq M cf_\ell^2(w)\right) \leq \exp\left(-\frac{1}{2} \frac{h^2 L}{M(1+h/3)}\right)$$

and

$$\mathbb{L} \max_{1 \leq \ell \leq L} \mathbb{P}(d_\ell(w) > \sqrt{M} cf_\ell(w) | E_{\ell,w}) \leq \mathbb{L} \delta^M.$$

Therefore,

$$\begin{aligned} \mathbb{P}\left(\exists w \in A^* : E_w, \frac{1}{L} \sum_{\ell=1}^L \frac{d_\ell^2(w)}{cf_\ell^2(w)} > 1 + h\right) &\leq \left(\exp\left(-\frac{1}{2} \frac{h^2 L}{M(1+h/3)}\right) + \mathbb{L} \delta^M\right) \sum_{w \in A^*} \mathbb{P}(E_w) \\ &\leq \frac{n^2}{2} \left(\exp\left(-\frac{1}{2} \frac{h^2 L}{M(1+h/3)}\right) + \mathbb{L} \delta^M\right). \end{aligned}$$

□

PROOF OF THEOREM 8, Good₂. Let $d_\ell(w) = d_\ell(\hat{p}_{n,\ell}(\cdot|w), \bar{p}_{n,\ell}(\cdot|w))$, $E_{\ell,w} = \{N_{n-1,\ell}(w) > 0\}$, and $E_w = \{\min_{1 \leq \ell \leq L} N_{n-1,\ell}(w) > 0\}$.

Let $M := \log(n^2 L / \delta) / \log \log n \geq 1$, $\mu = 1 / \log n$, $\delta_c = \mu / [2(1 + \mu)M|\mathcal{S}|]$, $\mu' := (3/2) \sqrt{\frac{\log(n^2 L / \delta) \cdot \log(n^2 / \delta)}{L \log \log n}}$ and $c_\ell(w)$ as defined (C.2) and (C.3).

By Lemma 5 applied to each $S \in \mathcal{S}$, we have that

$$\mathbb{P}(d_\ell(w) > c_\ell(w) | E_{\ell,w}) \leq 2|\mathcal{S}| \delta_c = \mu / [(1 + \mu)M].$$

By definition of μ , $c_\ell(w)$, and the relation above, we have

$$\begin{aligned} \mathbb{E}\left[\frac{d_\ell^2(w)}{(1+\mu)c_\ell^2(w)} \mid d_\ell^2(w) \leq M c_\ell^2(w), E_{\ell,w}\right] &\leq \mathbb{E}\left[\frac{d_\ell^2(w)}{(1+\mu)c_\ell^2(w)} \mid d_\ell^2(w) \leq c_\ell^2(w), E_{\ell,w}\right] + \\ &\quad + M \mathbb{P}(d_\ell(w) > c_\ell(w) | E_{\ell,w}) \\ &\leq \frac{1}{1+\mu} + M \cdot \frac{\mu}{(1+\mu)M} = 1. \end{aligned}$$

By Lemma 5 it also follows that for any $M \geq 1$ we have $P(d_\ell(w) > M\sqrt{(1+\mu)c_\ell(w)}|E_{\ell,w}) \leq \delta_c^M$ and both conditions of Lemma 2 holds for the confidence band being $\sqrt{(1+\mu)c_\ell(w)}$.

Therefore

$$\mathbb{P}\left(\exists w \in A^* : E_w, \frac{1}{L} \sum_{\ell=1}^L \frac{d_\ell^2(w)}{(1+\mu)c_\ell^2(w)} > 1 + \mu'\right) \leq \frac{n^2}{2} \left(\exp\left(-\frac{1}{2} \frac{\mu'^2 L}{M(1+\mu'/3)}\right) + L\delta_c^M \right).$$

Note that $n^2 L \delta_c^M \leq \delta$ provided that $M \geq \log(n^2 L / \delta) / \log(1/\delta_c)$. Since $2(1+\mu)M|\mathcal{S}| \geq 1$, $\delta_c \leq \mu = 1/\log n$, our choice of M satisfies this relation.

Next note that $n^2 \exp\left(-\frac{1}{2} \frac{\mu'^2 L}{M(1+\mu'/3)}\right) \leq \delta$, provided that $\frac{\mu'}{1+\mu'/3} \geq \sqrt{\frac{M \log(n^2/\delta)}{L}}$ which is satisfied by

$$\mu' = \frac{3}{2} \sqrt{\frac{\log(n^2 L / \delta) \log(n^2 / \delta)}{\log \log n \quad L}}$$

since we assumed $\log^2(n^2 L / \delta) < 3L \log \log n$.

□

APPENDIX D: PROOFS OF SECTION 5

PROOF OF LEMMA 1. For $x \in A_{-\infty}^{-1}$ and $u \in \mathcal{U}$, denote by $V_x = V(x)$ the true value function, $P_{u,x}$ denote the true transition probability (infinite) matrix, $\widehat{V}_x = \widehat{V}(\widehat{T}_n(x))$ the value function associated with the estimated transition probabilities $\widehat{P}_{u,x} = \widehat{P}_{u,\widehat{T}_n(x)}$.

Each value function is the fixed point of contraction mappings H, \widehat{H} on the functions $W : A_{-\infty}^{-1} \rightarrow \mathbb{R}$. Formally, the mappings

$$H(W)(x) = \max_{u \in \mathcal{U}} \{f(x_{-1}, u) + \beta P_{u,x} W\} \quad \text{and} \quad \widehat{H}(W)(x) = \max_{u \in \mathcal{U}} \{f(x_{-1}, u) + \beta \widehat{P}_{u,x} W\}$$

are contractions with modulus β by Blackwell's sufficient conditions.

Therefore we have

$$\begin{aligned} \|\widehat{V} - V\|_\infty &\leq \|\widehat{V} - \widehat{H}(V)\|_\infty + \|\widehat{H}(V) - V\|_\infty \\ &= \|\widehat{H}(\widehat{V}) - \widehat{H}(V)\|_\infty + \|\widehat{H}(V) - H(V)\|_\infty \\ &\leq \beta \|\widehat{V} - V\|_\infty + \|\widehat{H}(V) - H(V)\|_\infty. \end{aligned}$$

Thus, $\|\widehat{V} - V\|_\infty \leq \|\widehat{H}(V) - H(V)\|_\infty / (1 - \beta)$. where $\|\widehat{H}(V) - H(V)\|_\infty =$

$\max_{x \in A_{-\infty}^{-1}} |H(V)(x) - \widehat{H}(V)(x)|$. Thus the result follows by showing that

$$\begin{aligned} |H(V)(x) - \widehat{H}(V)(x)| &= \left| \max_{u \in \mathcal{U}} \{f(x_{-1}, u) + \beta \widehat{P}_{u,x} V\} - \max_{\tilde{u} \in \mathcal{U}} \{f(x_{-1}, \tilde{u}) + \beta P_{\tilde{u},x} V\} \right| \\ &\leq \beta \max_{u \in \mathcal{U}} |(\widehat{P}_{u,x} - P_{u,x})V| \\ &= \beta \max_{u \in \mathcal{U}} \left| \sum_{a \in A} [\hat{p}_{n,u}(a|\widehat{T}_n(x)) - p_u(a|x)] V(ax) \right| \\ &\leq \beta \|V(\cdot|x)\|_{\frac{q}{q-1}} \max_{u \in \mathcal{U}} \|\widehat{P}_{n,u}(\cdot|x) - p_u(\cdot|x)\|_q. \end{aligned}$$

□

PROOF OF THEOREM 9. The result follows by applying Lemma 1 with $q = 1$, which correspond to $d_\ell = \|\cdot\|_1/2$, and Theorem 5 with $r = k = m = \infty$ to bound

$$\max_{u \in \mathcal{U}} \frac{\|\widehat{P}_{u,x}(\cdot|x) - p_u(\cdot|x)\|_1}{2} \leq c_{T(x)} + \max \left\{ (1 + 2c) \|\text{cf}(T(x))\|_{L,\infty}, \frac{c+1}{c-1} c_{T(x)} \right\}.$$

The bound on $\|\text{cf}(T(x))\|_{L,\infty}$ follows from Theorem 6 with the family of sets $\mathcal{S} = 2^A$. □

PROOF OF THEOREM 10. By the choice of cf we have that with probability at least $1 - \delta$ the event Good_m occurs. Fix $a \in A$, $x, y \in A_{-\infty}^{-1}$ let

$$\bar{m}_\ell(a, x, y) = \bar{p}_{n,\ell}(a|T(x)) - \bar{p}_{n,\ell}(a|T(y)), \quad \overline{\text{AVEm}}(a, x, y) = \mathbb{E} [\bar{m}_\ell(a, x, y)],$$

and note that

(D.1)

$$\begin{aligned} \widehat{\text{AVEm}}(a, x, y) - \text{AVEm}(a, x, y) &= \frac{1}{L} \sum_{\ell=1}^L [\hat{m}_\ell(a, x, y) - \bar{m}_\ell(a, x, y)] + \\ &\quad + \frac{1}{L} \sum_{\ell=1}^L \bar{m}_\ell(a, x, y) - \overline{\text{AVEm}}(a, x, y) + \\ &\quad + \overline{\text{AVEm}}(a, x, y) - \text{AVEm}(a, x, y). \end{aligned}$$

First note that for $d_\ell = \|\cdot\|_\infty$, we have

$$\begin{aligned} |\overline{\text{AVEm}}(a, x, y) - \text{AVEm}(a, x, y)| &\leq \|d(\bar{p}_n(\cdot|T(x)), p(\cdot|x))\|_{L,1} + \|d(\bar{p}_n(\cdot|T(y)), p(\cdot|y))\|_{L,1} \\ &\leq c_{T(x)} + c_{T(y)}. \end{aligned}$$

Next, define $E_L(a, x, y) = \left| \frac{1}{L} \sum_{\ell=1}^L \bar{m}_\ell(a, x, y) - \overline{\text{AVEm}}(a, x, y) \right|$ and note that since $|\bar{m}_\ell(a, w, w')| \leq 1$ we have $E_L(a, x, y) \leq [2/L] + [(L-1)/L] E_{L-1}(a, x, y)$.

Moreover, note that $E_L(a, x, y) = E_L(a, T(x), T(y))$. Thus, we need to consider only suffixes $w \in T$ (in particular, leaves of T) which implies that $N_{n,\ell}(w) > 0$ for every $\ell = 1, \dots, L$. Thus, for $\xi = (\epsilon - [2/L])L/(L-1)$ we have

$$\begin{aligned} P\left(\max_{a \in A, w, w' \in T} E_L(a, w, w') \geq \epsilon\right) &\leq P\left(\max_{a \in A, w, w' \in T} E_{L-1}(a, w, w') \geq \xi\right) \\ &\leq P\left(\max_{\substack{a \in A, \\ w: \min_{\ell} N_{n,\ell}(w) > 0, \\ w': \min_{\ell} N_{n,\ell}(w') > 0}} E_{L-1}(a, w, w') \geq \xi\right) \\ &\leq P\left(\max_{\substack{a \in A, \\ w: N_{n,L}(w) > 0, \\ w': N_{n,L}(w') > 0}} E_{L-1}(a, w, w') \geq \xi\right) \\ &\leq \sum_{\substack{a \in A, \\ w \in A^*, \\ w' \in A^*}} P(N_{n,L}(w) > 0, N_{n,L}(w') > 0, E_{L-1}(a, w, w') \geq \xi). \end{aligned}$$

The event $\{N_{n,L}(w) > 0, N_{n,L}(w') > 0\}$ is independent of $\{E_{L-1}(a, w, w') \geq \xi\}$. Moreover, since $|\bar{m}_\ell(a, w, w')| \leq 1$ are i.i.d. draws from the population of agents, for any $\xi > 0$

$$P(E_{L-1}(a, w, w') > \xi) \leq \exp(-(L-1)\xi^2/2)$$

by Hoeffding's inequality. Therefore,

$$P\left(\max_{a \in A, w, w' \in T} E_L(a, w, w') \geq \epsilon\right) \leq |A| \exp(-(L-1)\xi^2/2) \cdot \sum_{\substack{w \in A^* \\ w' \in A^*}} P(N_{n,L}(w) > 0, N_{n,L}(w') > 0).$$

Next note that $\sum_{\substack{w \in A^* \\ w' \in A^*}} P(N_{n,L}(w) > 0, N_{n,L}(w') > 0)$ is the expected number of different pairs $w, w' \in A^*$ appearing in $X_1^n(\ell)$. Therefore we have

$$\sum_{\substack{w \in A^* \\ w' \in A^*}} P(N_{n,L}(w) > 0, N_{n,L}(w') > 0) \leq n^4/4.$$

Finally, to control the first term of (D.1), since $d_\ell = \|\cdot\|_\infty$, we have

$$\begin{aligned} \left|\frac{1}{L} \sum_{\ell=1}^L [\hat{m}_\ell(a, x, y) - \bar{m}_\ell(a, x, y)]\right| &\leq \left|\frac{1}{L} \sum_{\ell=1}^L [\hat{p}_{n,\ell}(a|\hat{T}_n(x)) - \bar{p}_{n,\ell}(a|T(x))]\right| + \\ &\quad + \left|\frac{1}{L} \sum_{\ell=1}^L [\hat{p}_{n,\ell}(a|\hat{T}_n(y)) - \bar{p}_{n,\ell}(a|T(y))]\right| \\ &\leq \left\|d(\hat{p}_n(\cdot|\hat{T}_n(x)), \bar{p}_n(\cdot|T(x)))\right\|_{L,1} + \\ &\quad + \left\|d(\hat{p}_n(\cdot|\hat{T}_n(y)), \bar{p}_n(\cdot|T(y)))\right\|_{L,1}. \end{aligned}$$

Under the event Good_2 , by Theorem 5, uniformly over $z \in A_{-\infty}^{-1}$ we have that

$$\left\|d(\hat{p}_n(\cdot|\hat{T}_n(z)), \bar{p}_n(\cdot|T(z)))\right\|_{L,1} \leq \max\left\{(1+2c)\|\text{cf}(T(z))\|_{L,2}, \frac{c+1}{c-1}c_{T(z)}\right\}$$

The result follows by combining these bounds. \square

APPENDIX E: A COMPENDIUM OF MARTINGALE RESULTS

LEMMA 3. *Let $(M_i, \mathcal{F}_i)_{i=0}^m$ be a martingale with $M_0 = 0$ and $|M_i - M_{i-1}| \leq Y_{i-1}$ for some \mathcal{F}_{i-1} -measurable r.v. Y_{i-1} . Define $V_n \equiv \sum_{j=0}^{n-1} Y_j^2$. Then:*

$$\forall \lambda, v > 0 : \mathbb{P}(M_n \geq \lambda, 0 < V_n \leq v) \leq \mathbb{P}(V_n > 0) e^{-\frac{\lambda^2}{2v}}.$$

PROOF. *Step 1: Main arguments.* Write $E \equiv \{\sum_{j=0}^{n-1} Y_j^2 > 0\}$ and define

$$U_r \equiv e^{sM_r - \frac{s^2}{2} \sum_{j=0}^{r-1} Y_j^2} \quad (0 \leq r \leq n)$$

where $s > 0$ will be fixed later. By Step 2 below (U_r, \mathcal{F}_r) is a supermartingale.

Now notice that:

$$M_n \geq \lambda, \sum_{j=0}^{n-1} Y_j^2 \leq v \Rightarrow sM_n - s\lambda + \frac{s^2 v}{2} - \frac{s^2}{2} \sum_{j=0}^{n-1} Y_j^2 \geq 0 \Rightarrow U_n e^{\frac{s^2 v}{2} - s\lambda} \geq 1.$$

Therefore,

$$\mathbb{P}\left(M_n \geq \lambda, 0 < \sum_{j=0}^{n-1} Y_j^2 \leq v\right) \leq e^{\frac{s^2 v}{2} - s\lambda} \mathbb{E}[U_n \chi_E].$$

The result follows by considering $s = \lambda/v$ and noting that $\mathbb{E}[U_n \chi_E] \leq \mathbb{P}(V_n > 0)$ by Step 3 below.

Step 2: (U_r, \mathcal{F}_r) is a supermartingale. Since Y_r is \mathcal{F}_r -measurable,

$$\mathbb{E}\left[\frac{U_{r+1}}{U_r} \mid \mathcal{F}_r\right] = \mathbb{E}\left[e^{s(M_{r+1} - M_r)} \mid \mathcal{F}_r\right] e^{-\frac{s^2 Y_r^2}{2}}.$$

Recall that $|M_{r+1} - M_r| \leq Y_r$, hence by convexity

$$e^{s(M_{r+1} - M_r)} \leq \cosh(sY_r) + \sin(sY_r)(M_{r+1} - M_r).$$

Taking conditional expectations, we see that:

$$\mathbb{E}\left[e^{s(M_{r+1} - M_r)} \mid \mathcal{F}_r\right] \leq \cosh(sY_r) + \sin(sY_r) \mathbb{E}[(M_{r+1} - M_r) \mid \mathcal{F}_r] = \cosh(sY_r).$$

This implies $\mathbb{E}\left[e^{s(M_{r+1} - M_r)} \mid \mathcal{F}_r\right] e^{-\frac{s^2 Y_r^2}{2}} \leq \cosh(sY_r) e^{-\frac{s^2 Y_r^2}{2}} \leq 1$ since $\cosh(x) \leq e^{x^2/2}$ for all $x \in \mathbb{R}$.

Step 3: $\mathbb{E}[U_n \chi_E] \leq \mathbb{P}(V_n > 0)$. Write $E_0 \equiv \{Y_0 \neq 0\}$ and $E_j \equiv \{Y_j \neq 0\} \cap \{Y_k = 0, 0 \leq k < j\}$. Notice that $E = \cup_{0 \leq j \leq n-1} E_j$ (where the union is

disjoint) and that each E_j is \mathcal{F}_j -measurable. Moreover, if E_k holds, we have $\sum_{j=0}^{k-1} Y_j^2 = 0$ and $M_k = M_0 = 0$, hence $U_k = 1$. Therefore,

$$\begin{aligned} \mathbb{E}[U_n \chi_E] &= \sum_{k=0}^{n-1} \mathbb{E}[U_n \chi_{E_k}] \\ (E_k \text{ is } \mathcal{F}_k\text{-measurable}) &= \sum_{k=0}^{n-1} \mathbb{E}[\chi_{E_k} \mathbb{E}[U_n | \mathcal{F}_k]] \\ (U_k \text{ is supermartingale}) &\leq \sum_{k=0}^{n-1} \mathbb{E}[\chi_{E_k} U_k] \\ (U_k = 1 \text{ in } E_k) &= \sum_k \mathbb{P}(E_k) = \mathbb{P}(E). \end{aligned}$$

□

LEMMA 4. *Let $(M_i, \mathcal{F}_i)_{i=0}^n$ be a martingale with $M_0 = 0$, $|M_i - M_{i-1}| \leq Y_{i-1} \leq 1$ for some \mathcal{F}_{i-1} -measurable r.v. Y_{i-1} . Define $V_n \equiv \sum_{j=1}^n E[(M_j - M_{j-1})^2 | \mathcal{F}_{j-1}]$. Then:*

$$\forall \lambda, v > 0 : \mathbb{P}(M_n \geq \lambda, 0 < V_n \leq v) \leq \mathbb{P}(V_n > 0) e^{-\frac{\lambda^2}{2v}(2 - \exp(\lambda/v))}.$$

PROOF. *Step 1: Main Arguments.* Write $E \equiv \{V_n > 0\}$ and define

$$U_r \equiv e^{sM_r - \frac{s^2 e^s}{2} \sum_{j=1}^r E[(M_j - M_{j-1})^2 | \mathcal{F}_{j-1}]} \quad (0 \leq r \leq n)$$

where $s > 0$ will be fixed later. It follows that (U_r, \mathcal{F}_r) is a supermartingale by Step 2 below.

Now notice that:

$$M_n \geq \lambda, V_n \leq v \Rightarrow sM_n - s\lambda + \frac{s^2 e^s v}{2} - \frac{s^2 e^s}{2} V_n \geq 0 \Rightarrow U_n e^{\frac{s^2 e^s v}{2} - s\lambda} \geq 1.$$

Therefore,

$$\begin{aligned} \mathbb{P}(M_n \geq \lambda, 0 < V_n \leq v) &= \mathbb{E} \left[\mathbb{1}_{\{U_n e^{\frac{s^2 e^s v}{2} - s\lambda} \geq 1\}} \cdot \chi_E \right] \\ &\leq \mathbb{E} \left[U_n e^{\frac{s^2 e^s v}{2} - s\lambda} \chi_E \right] \\ &= e^{\frac{s^2 e^s v}{2} - s\lambda} \mathbb{E}[U_n \chi_E]. \end{aligned}$$

The Lemma follows from choosing $s = \lambda/v$ and noting that $\mathbb{E}[U_n \chi_E] \leq \mathbb{P}(E)$ by Step 3.

Step 2: (U_r, \mathcal{F}_r) is a supermartingale. Since Y_r is \mathcal{F}_r -measurable,

$$\mathbb{E} \left[\frac{U_{r+1}}{U_r} \mid \mathcal{F}_r \right] = \mathbb{E} \left[e^{s(M_{r+1}-M_r)} \mid \mathcal{F}_r \right] e^{-\frac{s^2 e^s E[(M_{r+1}-M_r)^2 \mid \mathcal{F}_r]}{2}}.$$

Recall that $|M_{r+1} - M_r| \leq Y_r \leq 1$, hence by [20] page 32,

$$\mathbb{E} \left[e^{s(M_{r+1}-M_r)} \mid \mathcal{F}_r \right] \leq \exp \left(\frac{s^2 e^s}{2} E[(M_{r+1} - M_r)^2 \mid \mathcal{F}_r] \right).$$

This implies $\mathbb{E} \left[e^{s(M_{r+1}-M_r)} \mid \mathcal{F}_r \right] e^{-\frac{s^2 e^s E[(M_{r+1}-M_r)^2 \mid \mathcal{F}_r]}{2}} \leq 1$.

Step 3: $\mathbb{E}[U_n \chi_E] \leq \mathbb{P}(E)$. The proof is similar to Step 3 in Lemma 3. \square

For any $\gamma > 1$, $\delta \in (0, 1)$, define monotonic function $h : [0, \infty) \rightarrow [0, \infty)$

$$h(x) = 2x\gamma^2 \log \left\{ \frac{2}{\delta} (1 + \log_\gamma x)(2 + \log_\gamma x) \right\},$$

and let i_0 be the smallest integer such that

$$\gamma^{i_0} \log^2(2 - (1/\gamma)) \geq 2 \log(2/\delta) + 2 \log[(1 + i_0)(2 + i_0)],$$

so that $2 - \exp(\sqrt{h(\gamma^{i_0})}/\gamma^{i_0+1}) \geq 1/\gamma$.

LEMMA 5. *Let $(M_i, \mathcal{F}_i)_{i=0}^m$ be a martingale with $M_0 = 0$, $|M_i - M_{i-1}| \leq Y_{i-1} \leq 1$ for some \mathcal{F}_{i-1} -measurable binary r.v. Y_{i-1} . Define $V_n \equiv \sum_{j=1}^n E[(M_j - M_{j-1})^2 \mid \mathcal{F}_{j-1}]$ and $\tilde{V}_n \equiv \sum_{j=0}^{n-1} Y_j^2$. Then*

$$\mathbb{P} \left(M_n \geq \sqrt{h(V_n)}, V_n > \gamma^{i_0} \right) + \mathbb{P} \left(M_n \geq \sqrt{h(\tilde{V}_n)/\gamma}, 0 < V_n \leq \gamma^{i_0} \right) \leq \delta \cdot \mathbb{P}(V_n > 0).$$

PROOF. First note that

$$\begin{aligned} P \left(\sum_{i=1}^n d_i \geq \sqrt{h(V_n)}, V_n > \gamma^{i_0} \right) &= \sum_{i \geq i_0} P \left(\sum_{i=1}^n d_i \geq \sqrt{h(V_n)}, \gamma^i \leq V_n \leq \gamma^{i+1} \right) \\ &\leq \sum_{i \geq i_0} P \left(\sum_{i=1}^n d_i \geq \sqrt{h(\gamma^i)}, 0 < V_n \leq \gamma^{i+1} \right) \\ &\leq P(V_n > 0) \sum_{i \geq i_0} \exp \left(-\frac{h(\gamma^i)}{2\gamma^{i+1}} \left[2 - \exp \left(\frac{\sqrt{h(\gamma^i)}}{\gamma^{i+1}} \right) \right] \right) \\ &\leq P(V_n > 0) \sum_{i \geq i_0} \exp \left(-\log \frac{2}{\delta} - \log[(1+i)(2+i)] \right) \end{aligned}$$

where the last line follows from the definition of i_0 . Since $i_0 \geq 0$ it follows that

$$\sum_{i \geq i_0} \exp\left(-\log \frac{2}{\delta} - \log[(1+i)(2+i)]\right) = \frac{\delta}{2} \sum_{i \geq i_0} \frac{1}{(1+i)(2+i)} \leq \frac{\delta}{2}.$$

Next, note that the event $0 < V_n < \gamma^{i_0} \subseteq \tilde{V}_n > 0$ and that \tilde{V}_n only takes integer values, hence:

$$\{\tilde{V}_n > 0\} = \bigcup_{i=0}^{+\infty} E_i \text{ where } E_i \equiv \{\gamma^i \leq \tilde{V}_n < \gamma^{i+1}\}$$

and the union is disjoint. We deduce

$$\begin{aligned} P\left(\sum_{i=1}^n d_i \geq \sqrt{h(\tilde{V}_n)/\gamma}, \tilde{V}_n > 0\right) &= \sum_{i \geq 0} P\left(\sum_{i=1}^n d_i \geq \sqrt{h(\tilde{V}_n)/\gamma}, \gamma^i \leq \tilde{V}_n < \gamma^{i+1}\right) \\ &\leq \sum_{i \geq 0} P\left(\sum_{i=1}^n d_i \geq \sqrt{h(\gamma^i)/\gamma}, 0 < \tilde{V}_n < \gamma^{i+1}\right) \\ &\leq P(V_n > 0) \sum_{i \geq 0} \exp\left(-\frac{h(\gamma^i)}{2\gamma^{i+2}}\right) \\ &\leq P(V_n > 0) \sum_{i \geq i_0} \exp\left(-\log \frac{2}{\delta} - \log[(1+i)(2+i)]\right) \\ &\leq P(V_n > 0) \delta/2. \end{aligned}$$

□

APPENDIX F: PROOFS OF SECTION 6

PROOF OF THEOREM 11. Fix $\tilde{T} \in \mathcal{T}_{\delta_0}$. Since \tilde{T} is complete and finite, $\tilde{T}(x)$ is always a leaf of T . Thus our goal can be restated as:

$$(F.1) \quad \sup_{w \in S} \left| \frac{\|\mathbf{cf}(w)\|_{L,r}}{\|\overline{\mathbf{cf}}(w)\|_{L,r}} - 1 \right| \leq \frac{C_2}{\log n},$$

where S is the set of leaves of \tilde{T} . To prove this, we apply Theorem 17 to each of the processes $X(\ell)_{-\infty}^{+\infty}$, replacing n with $n-1$ and choosing the other parameters as follows:

$$\delta = \delta_* \equiv \frac{\delta_0}{L}, \quad \xi = \frac{1}{\ln n}, \quad S = \{w \in A^* : w \text{ is a leaf of } \tilde{T}, \min_{1 \leq \ell \leq L} \pi_\ell(w) > 0\}.$$

Notice that the quantities π_S and h_S in Theorem 17 equal $\pi_{\tilde{T}}$ and $h_{\tilde{T}}$ (resp.). Moreover, $\pi_1(w) \geq \pi_{\tilde{T}}$ for all $w \in S$ so that

$$|S| \pi_{\tilde{T}} \leq \sum_{w \in S} \mathbb{P} \left(\tilde{T}(X(1)_{-\infty}^{-1}) = w \right) \leq \sum_{w \in A^*} \mathbb{P} \left(\tilde{T}(X(1)_{-\infty}^{-1}) = w \right) \leq 1.$$

Hence $|S| \leq \pi_{\tilde{T}}^{-1} \leq n$ by our assumption on \tilde{T} . Using Remark 9 (after the Theorem 17), we see that its conclusion will hold whenever:

$$n \geq C_1 (1 + \alpha) \frac{(h_{\tilde{T}} + 1) \log^4 n}{\pi_{\tilde{T}}} \text{ and } n \geq C_1 (1 + \alpha) \frac{\log^3 n}{\pi_{\tilde{T}}} \beta^{-1} \left(\frac{\delta \pi_{\tilde{T}} \xi}{6} \right)$$

for some universal constant C_1 . The first condition holds since $\tilde{T} \in \mathcal{T}_{\delta_0}$. For the second, we notice that Assumption 2 implies $\beta^{-1}(x) \leq \left\lceil (\Gamma/x)^{\frac{1}{\gamma}} \right\rceil \leq 1 + (\Gamma/x)^{\frac{1}{\gamma}}$. Thus

$$\beta^{-1} \left(\frac{\delta_* \pi_{\tilde{T}} \xi}{6} \right) \leq 1 + (6\Gamma)^{\frac{1}{\gamma}} \delta_*^{-\frac{1}{\gamma}} \log^{\frac{1}{\gamma}} n \pi_{\tilde{T}}^{-\frac{1}{\gamma}} \leq \left(1 + (6\Gamma)^{\frac{1}{\gamma}} \right) \delta_*^{-\frac{1}{\gamma}} \log^{\frac{1}{\gamma}} n \pi_{\tilde{T}}^{-\frac{1}{\gamma}}.$$

Hence we need: $n \geq C (1 + \alpha) (1 + (6\Gamma)^{\frac{1}{\gamma}}) \delta_*^{-\frac{1}{\gamma}} \log^{\frac{3\gamma+1}{\gamma}} n \pi_{\tilde{T}}^{\frac{\gamma+1}{\gamma}}$, or equivalently

$$(F.2) \quad \pi_{\tilde{T}}^{-1} \leq \left(\frac{1}{[C_1 (1 + \alpha) (1 + (6\Gamma)^{\frac{1}{\gamma}})]^{\frac{\gamma}{\gamma+1}}} \right) \frac{\delta_*^{\frac{1}{\gamma+1}} n^{\frac{\gamma}{\gamma+1}}}{\log^{\frac{3\gamma+1}{\gamma+1}} n}$$

for some constant C_1 that is entirely determined by C . Since $\delta_* = \delta_0/L = n^{-\alpha_L} \delta_0$, (F.2) is also guaranteed by $\tilde{T} \in \mathcal{T}_{\delta_0}$ (if C_1 is chosen appropriately). We conclude that $\forall 1 \leq \ell \leq L$:

$$\mathbb{P} \left(\forall w \text{ leaf of } \tilde{T} : \left| \frac{N_{n-1, \ell}(w)}{\pi_{\ell}(w)(n-1)} - 1 \right| \leq \frac{1}{\log n} \right) \geq 1 - \delta_* = 1 - \frac{\delta_0}{L}.$$

Therefore,

$$\mathbb{P} \left(\forall 1 \leq \ell \leq L, \forall w \text{ leaf of } \tilde{T} : \left| \frac{N_{n-1, \ell}(w)}{\pi_{\ell}(w)(n-1)} - 1 \right| \leq \frac{1}{\log n} \right) \geq 1 - \delta_0.$$

Given that this last event occurs, direct calculations yield

$$\forall 1 \leq \ell \leq L, \forall w \in S, 1 - \frac{C_2}{\log n} \leq \frac{\text{cf}_{\ell}(w)}{\text{cf}_{\ell}(w)} \leq 1 + \frac{C_2}{\log n}$$

where $C_2 > 0$ is universal. From this one concludes (F.1) for any $r \geq 1$. \square

PROOF OF THEOREM 12. Theorem 5 shows that the occurrence of Good_m implies that for all $x \in A_{-\infty}^{-1}$

$$\left\| d(p(\cdot|x), \hat{P}_n(\cdot|x)) \right\|_{L,k} \leq c_{T^{\delta_0}(x)} + \max \left\{ \frac{c+1}{c-1} c_{T^{\delta_0}(x)}, (1+2c) \left\| \text{cf}(T^{\delta_0}(x)) \right\|_{L,r} \right\}.$$

However, $c_{T^{\delta_0}(x)} \leq \bar{c}_{T^{\delta_0}(x)}$ (by definition of the two quantities) and by the typicality event:

$$\left\| \text{cf}(T^{\delta_0}(x)) \right\|_{L,r} \leq \left(1 + \frac{C_2}{\log n} \right) \left\| \bar{\text{cf}}(T^{\delta_0}(x)) \right\|_{L,r}.$$

Thus, for all $x \in A_{-\infty}^{-1}$

$$\begin{aligned} \left\| d(p(\cdot|x), \hat{P}_n(\cdot|x)) \right\|_{L,k} &\leq c_{T^{\delta_0}(x)} + \max \left\{ \frac{c+1}{c-1} c_{T^{\delta_0}(x)}, \right. \\ &\quad \left. \left(1 + \frac{C_2}{\log n} \right) (1+2c) \left\| \bar{\text{cf}}(T^{\delta_0}(x)) \right\|_{L,r} \right\}. \end{aligned}$$

□

PROOF OF THEOREM 13. Let $w \in \tilde{T}_-$. We will show that the occurrence of $\text{Typ}_r(\tilde{T})$ implies that $\text{CanRmv}(w) = 0$, so that $w \in \tilde{T}_n$. By assumption, there exist $x \in A_{-\infty}^{-1}$ and $y \in A_{-\infty}^{-1}$ with $\tilde{T}(x) \succeq w$, $\tilde{T}(y) \succeq \text{par}(w)$ with:

(F.3)

$$\begin{aligned} \left\| d(p(\cdot|x), p(\cdot|y)) \right\|_{L,k} &> \left(1 + \frac{C_2}{\log n} \right) (1+c) \left(\left\| \bar{\text{cf}}(\tilde{T}(x)) \right\|_{L,r} + \left\| \bar{\text{cf}}(\tilde{T}(y)) \right\|_{L,r} \right) \\ &\quad + \bar{c}_{\tilde{T}(x)} + \bar{c}_{\tilde{T}(y)}. \end{aligned}$$

Notice that if $\text{Good}_m \cap \text{Typ}_r(\tilde{T})$ holds,

$$\begin{aligned} \left\| d(p(\cdot|x), \hat{p}_n(\cdot|\tilde{T}(x))) \right\|_{L,k} &\leq \left\| d(p(\cdot|x), \bar{p}_n(\cdot|\tilde{T}(x))) \right\|_{L,k} + \left\| d(\bar{p}_n(\cdot|\tilde{T}(x)), \hat{p}_n(\cdot|\tilde{T}(x))) \right\|_{L,k} \\ \text{(defn. of } \bar{c}_{\tilde{T}(x)} \text{)} &\leq \bar{c}_{\tilde{T}(x)} + \left\| d(\bar{p}_n(\cdot|\tilde{T}(x)), \hat{p}_n(\cdot|\tilde{T}(x))) \right\|_{L,k} \\ \text{(Good}_m \text{ holds and Thm. 1)} &\leq \bar{c}_{\tilde{T}(x)} + \left\| \text{cf}(T(x)) \right\|_{L,r} \\ \text{(Typ}_r(\tilde{T}) \text{ holds)} &\leq \bar{c}_{\tilde{T}(x)} + \left(1 + \frac{C_2}{\log n} \right) c \left\| \bar{\text{cf}}(\tilde{T}(x)) \right\|_{L,r}. \end{aligned}$$

Similarly, if $\text{Good}_m \cap \text{Typ}_r(\tilde{T})$ holds,

$$\left\| d(p(\cdot|y), \hat{p}_n(\cdot|\tilde{T}(y))) \right\|_{L,k} \leq \bar{c}_{\tilde{T}(y)} + \left(1 + \frac{C_2}{\log n} \right) \left\| \bar{\text{cf}}(\tilde{T}(y)) \right\|_{L,k}.$$

Combining these two inequalities shows that:

$$\begin{aligned} \left\| d(\hat{p}_n(\cdot|\tilde{T}(x)), \hat{p}_n(\cdot|\tilde{T}(y))) \right\|_{L,k} &\geq \|d(p(\cdot|y), p(\cdot|x))\|_{L,k} - \left\| d(p(\cdot|y), \hat{p}_n(\cdot|\tilde{T}(y))) \right\|_{L,k} - \left\| d(p(\cdot|x), \hat{p}_n(\cdot|\tilde{T}(x))) \right\|_{L,k} \\ &\geq d(p(\cdot|y), p(\cdot|x)) - \bar{c}_{\tilde{T}(x)} - \bar{c}_{\tilde{T}(y)} \\ &\quad - \left(1 + \frac{C_2}{\log n}\right) \left(\left\| \bar{\text{cf}}(\tilde{T}(x)) \right\|_{L,r} + \left\| \bar{\text{cf}}(\tilde{T}(y)) \right\|_{L,r} \right). \end{aligned}$$

We may now plug (F.3) to deduce that if $\text{Good}_m \cap \text{Typ}_r(\tilde{T})$ holds, then:

$$\begin{aligned} \left\| d(\hat{p}_n(\cdot|\tilde{T}(x)), \hat{p}_n(\cdot|\tilde{T}(y))) \right\|_{L,k} &\geq \|d(p(\cdot|y), p(\cdot|x))\|_{L,k} - \left\| d(p(\cdot|y), \hat{p}_n(\cdot|\tilde{T}(y))) \right\|_{L,k} - \left\| d(p(\cdot|x), \hat{p}_n(\cdot|\tilde{T}(x))) \right\|_{L,k} \\ &> \left(1 + \frac{C_2}{\log n}\right) c \left(\left\| \bar{\text{cf}}(\tilde{T}(x)) \right\|_{L,k} + \left\| \bar{\text{cf}}(\tilde{T}(y)) \right\|_{L,k} \right) \\ (\text{Typ}_r(\tilde{T}) \text{ holds}) &> c \left(\left\| \text{cf}(\tilde{T}(x)) \right\|_{L,k} + \left\| \text{cf}(\tilde{T}(y)) \right\|_{L,k} \right). \end{aligned}$$

Therefore, the occurrence of $\text{Good}_m \cap \text{Typ}_r(\tilde{T})$ implies that $\exists w' = \tilde{T}(x) \succeq w, w'' = \tilde{T}(y) \succeq \text{par}(w)$:

$$\left\| d(\hat{p}_n(\cdot|w'), \hat{p}_n(\cdot|w'')) \right\|_{L,k} > c \left(\left\| \text{cf}(w') \right\|_{L,k} + \left\| \text{cf}(w'') \right\|_{L,k} \right),$$

which implies $\text{CanRmv}(w) = 0$. \square

PROOF OF THEOREM 14. Fix some $x \in A_{-\infty}^{-1}$. If Good_m holds, for all $w, w' \succeq \tilde{T}_+(x)$,

$$\begin{aligned} d(\hat{p}_n(\cdot|w), \hat{p}_n(\cdot|w')) &\leq d(\bar{p}_n(\cdot|w), \bar{p}_n(\cdot|w')) + \left(\left\| \text{cf}(w) \right\|_{L,r} + \left\| \text{cf}(w') \right\|_{L,r} \right) \\ &\leq \bar{c}_{\tilde{T}_+(x)} + \left(\left\| \text{cf}(w) \right\|_{L,r} + \left\| \text{cf}(w') \right\|_{L,k} \right), \end{aligned}$$

since both $\bar{p}_n(\cdot|w)$ and $\bar{p}_n(\cdot|w')$ are convex combinations of probabilities of the form $p(\cdot|y)$ with $y \succeq \tilde{T}_+(x)$. We have also assumed:

$$\bar{c}_{\tilde{T}_+(x)} \leq 2 \left(1 - \frac{C_2}{\log n}\right) (c-1) \left\| \bar{\text{cf}}(\tilde{T}_+(x)) \right\|_{L,r},$$

and if $\text{Typ}_r(\tilde{T}_+)$ holds, we have

$$\left(1 - \frac{C_2}{\log n}\right) \left\| \bar{\text{cf}}(\tilde{T}_+(x)) \right\|_{L,r} \leq \left\| \text{cf}(\tilde{T}_+(x)) \right\|_{L,r}.$$

Hence under $\text{Good}_m \cap \text{Typ}_r(\tilde{T}_+)$: for all $w, w' \succeq \tilde{T}_+(x)$,

$$\begin{aligned} d(\hat{p}_n(\cdot|w), \hat{p}_n(\cdot|w')) &\leq 2(c-1) \left\| \text{cf}(\tilde{T}_+(x)) \right\|_{L,r} + \left(\left\| \text{cf}(w) \right\|_{L,r} + \left\| \text{cf}(w') \right\|_{L,r} \right) \\ &\leq c \left(\left\| \text{cf}(w) \right\|_{L,r} + \left\| \text{cf}(w') \right\|_{L,r} \right), \end{aligned}$$

since the confidence radii are monotone. This last inequality implies that:

$$\forall w \succ \tilde{T}_+(x), w_1 \succeq w, w_2 \succeq \text{par}(w),$$

$$d(\hat{p}_n(\cdot|w_1), \hat{p}_n(\cdot|w_2)) \leq c \left(\|\text{cf}(w_1)\|_{L,r} + \|\text{cf}(w_2)\|_{L,r} \right).$$

In particular, all $w \succ \tilde{T}_+(x)$ are such that $\text{CanRmv}(w) = 1$. But any w that does *not* belong to \tilde{T}_+ satisfies $w \succ \tilde{T}_+(x)$ for some x . It follows that all nodes v with $\text{CanRmv}(v) = 0$ belong to \tilde{T}_+ . In particular, $\hat{T}_n \subseteq \tilde{T}_+$. \square

PROOF OF THEOREM 15. Take $\gamma = (\alpha + 1) \log n$, $\Gamma_\gamma \equiv \chi (\gamma/[e\nu])^\gamma$. Simple calculus shows that the function $\beta(\cdot)$ appearing in the assumption satisfies: $\forall b \in \mathbb{N}$, $\beta(b) \leq \Gamma_\gamma b^{-\gamma}$. Hence Assumption 2 is satisfied.

We now *claim* that:

CLAIM 3. *If (6.3) holds, $T^* \in \mathcal{T}_\delta$.*

PROOF OF THE CLAIM. We need to check two conditions:

$$(F.4) \quad \pi_{T^*}^{-1} \leq \left(\frac{1}{[C_1(1+\alpha)(1+(6\Gamma)^{\frac{1}{\gamma}})]^{\frac{\gamma}{\gamma+1}}} \right) \frac{\left(\frac{\delta}{L}\right)^{\frac{1}{\gamma+1}} n^{\frac{\gamma}{\gamma+1}}}{\log^{\frac{3\gamma+1}{\gamma+1}} n};$$

$$(F.5) \quad h_{T^*} + 1 \leq \frac{\pi_{T^*} n}{C_1(1+\alpha) \log^4 n}.$$

Equation (F.5) is clearly a consequence of (6.3), at least if $C_3 \geq C_1$. To check (F.4), we note that, at the cost of replacing C_1 by $C_1 \vee 1$, we may assume that the denominator in the bracketed fraction in the RHS is at least 1. In this case we may replace $\gamma/(\gamma+1)$ by 1 while decreasing the RHS. Moreover,

$$(6\Gamma)^{\frac{1}{\gamma}} = (6\chi)^{1/(\alpha+1) \log n} \frac{(\alpha+1) \log n}{e\nu} \leq \frac{(\alpha+1) \log n}{\nu}$$

since $6\chi \leq n^{\alpha+1}$. We deduce:

$$\left(\frac{1}{[C_1(1+\alpha)(1+(6\Gamma)^{\frac{1}{\gamma}})]^{\frac{\gamma}{\gamma+1}}} \right) \geq \left(\frac{1}{C_1(1+\alpha)(1+e^2 \left(\frac{\alpha+1}{\nu}\right) \log n)} \right).$$

We also note that $n \geq 9 \Rightarrow \log^{\frac{3\gamma+1}{\gamma+1}} n \leq \log^3 n$ and (since $L/\delta = n^\alpha$)

$$(\delta/L)^{\frac{1}{\gamma+1}} n^{\frac{\gamma}{\gamma+1}} = n^{\frac{\gamma-\alpha}{\gamma+1}} \geq n^{1-\frac{\alpha+1}{\gamma}} \geq n/e$$

by the choice of $\gamma = (\alpha + 1) \log n$. We conclude that (F.4) is also implied by (F.5) in the Theorem if C_3/C_1 is large enough, which we can ensure by the appropriate choice of C_3 . \square

Combining the Claim with the definition of \mathcal{T}_{δ_0} , by Theorem 11 we have $\mathbb{P}(\text{Typ}_r(T^*)) \geq 1 - \delta_0$. Now assume that $\text{Typ}_r(T^*) \cap \text{Good}_m$ occurs, which will be the case with probability $1 - \delta - \delta_0$. In that case the bound on $\left\|d(p(\cdot|x), \widehat{P}_n(\cdot|x))\right\|_{L,k}$ follows from combining Theorem 12 and the fact that $\bar{c}_{T^*(x)} = 0$ under the parametric case (Assumption 3). We also have $\widehat{T}_n \subseteq T^*$ from Theorem 2.

To finish, we must also prove that $\widehat{T}_n \supseteq T^*$ via Theorem 13 applied to $\widetilde{T} = \widetilde{T}_- = T^*$. In order to check the assumptions, we note that, under Assumption 3, if $T^*(x)$ is a leaf of T^* , then $\bar{c}_{T^*(x)} = 0$. By definition of $d_{k,\min}$, we also have

$$\|d(p(\cdot|x), p(\cdot|y))\|_{L,k} = \|d(p(\cdot|T^*(x)), p(\cdot|T^*(y)))\|_{L,k} \geq d_{k,\min}$$

for any $y \in A_{-\infty}^{-1}$ with $T^*(y) \succeq \text{par}(T^*(x))$. Hence the condition in Theorem 13 is implied by:

$$2 \left(1 + \frac{C_2}{\log n}\right) (1 + c) \max_{x \in A_{-\infty}^{-1}} \left\| \overline{\text{cf}}(T^*(x)) \right\|_{L,r} < d_{k,\min},$$

where the minimum can be taken over the (finite many) leaves of T^* instead. \square

PROOF OF THEOREM 16. Comets, Fernández and Ferrari [10] have shown that processes with continuity rates $\lambda(b) = O(b^{-1-\gamma})$ and $q_{\min} > 0$ are φ -mixing with rate function $O(b^{-\gamma})$. Since φ -mixing implies β -mixing with the same rate function, we deduce from Assumption 4 that the processes $X(1), \dots, X(L)$ are β -mixing with common rate function $\beta(b) \equiv \Gamma b^{-\gamma}$ ($b \in \mathbb{N}$) where $\Gamma > 0$ depends on q_{\min} and Γ_0 . This brings us into the realm of Assumption 2. We also note for later use that our assumption on q_{\min} implies:

$$(F.6) \quad \forall w \in A^*, \forall 1 \leq \ell \leq L : \pi_\ell(w) \geq q_{\min}^{|w|}.$$

Now let $h \equiv \lceil \log n / [3 \log(1/q_{\min})] \rceil$ and define $T_h \subset A^*$ as the tree containing all strings of length $\leq h$. Our next goal is to prove the following Claim.

CLAIM 4. *Let $\delta_0 \equiv L n^{-\frac{\gamma-1}{2}}$. Under the assumptions of the Theorem, $T_h \in \mathcal{T}_{\delta_0}$.*

PROOF OF THE CLAIM. We need to show that:

$$(F.7) \quad \pi_{T_h}^{-1} \leq \left(\frac{1}{[C_1(1+\alpha)(1+(6\Gamma)^{\frac{1}{\gamma}})]^{\frac{\gamma}{\gamma+1}}} \right) \frac{\left(\frac{\delta_0}{L}\right)^{\frac{1}{\gamma+1}} n^{\frac{\gamma}{\gamma+1}}}{\log^{\frac{3\gamma+1}{\gamma+1}} n};$$

$$(F.8) \quad h_{T_h} + 1 \leq \frac{\pi_{T_h} n}{C_1(1+\alpha) \log^4 n}.$$

To prove (F.7), we note that $3\gamma + 1 \leq 3(\gamma + 1)$ and, by definition of δ_0 , $(\delta_0/L)^{\frac{1}{\gamma+1}} n^{\frac{\gamma}{\gamma+1}} = \sqrt{n}$. Hence we may underestimate the RHS of (F.7) by:

$$\frac{\sqrt{n}}{C_6(1+\alpha) \log^3 n}$$

where C_6 depends on Γ and γ only. On the other hand, (F.6) applied to the leaves of T_h gives:

$$(F.9) \quad \pi_{T_h}^{-1} \leq \left(\frac{1}{q_{\min}}\right)^h \leq \left(\frac{1}{q_{\min}}\right) n^{1/3}.$$

The inequality (F.7) thus follows from:

$$\left(\frac{1}{q_{\min}}\right) n^{1/3} \leq \frac{\sqrt{n}}{C_6(1+\alpha) \log^3 n}$$

which is implied by (6.4) if C_4 is large enough (this will only depend on Γ , γ and q_{\min}).

We use the same upper bound on $\pi_{T_h}^{-1}$ to check (F.8). Thus, the RHS can be lower bounded by:

$$\frac{n^{2/3}}{\left(\frac{1}{q_{\min}}\right) \log^4 n},$$

which is larger than h if (6.4) holds (if C_4 is large enough). \square

We may now use T_h to upper bound the estimation error appearing in Theorem 12 since T^{δ_0} yields a bound at least as good as T_h . We first need bounds on $\bar{c}_{T_h(x)}$. We have noted before the statement of Theorem 16 that the total variation distance between $\tilde{p}_\ell(\cdot|T_b(x))$ and $p_\ell(a|x)$ is at most $\lambda_\ell(b)$. By the triangle inequality,

$$\forall 1 \leq \ell \leq L, \bar{c}_{T_h(x)} \leq \left\| \{2\lambda_\ell(h)\}_{\ell=1}^L \right\|_{L,k} \leq 2\Gamma_0 h^{-\gamma-1} \leq \frac{C_5}{\log^{\gamma+1} n}$$

with C_5 as in the Theorem. On the other hand, since for all leaves w of T_h we have

$$\pi_\ell(w) n \geq \pi_{T_h} n \geq \frac{n^{2/3}}{\log\left(\frac{1}{q_{\min}}\right)} \geq \frac{(C_4)^{2/3} (1 + \alpha)^4 \log^{2(\gamma+2)} n}{\log\left(\frac{1}{q_{\min}}\right)}.$$

One may use this to check that

$$\sup_{x \in A_{-\infty}^{-1}} \left\| \overline{\text{cf}}(T_h(x)) \right\|_{L,r} = O\left(\frac{1}{\log^{\gamma+1} n}\right)$$

where the asymptotics are for q_{\min}, γ and Γ_0 fixed and n large. Increasing C_5 if necessary, we conclude that for all $x \in A_{-\infty}^{-1}$:

$$\bar{c}_{T_h(x)} + \max\left\{\frac{c+1}{c-1}\bar{c}_{T_h(x)}, (1+2c)\left\|\overline{\text{cf}}(T_h(x))\right\|_{L,r}\right\} \leq \frac{C_5}{\log^{\gamma+1} n},$$

and this implies the desired result via Theorem 12 and some further adjustments of C_5 . \square

APPENDIX G: TYPICALITY RESULTS FOR β -MIXING PROCESSES

In what follows we use

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

THEOREM 17. *Let $X_{-\infty}^{+\infty}$ be a stationary and β -mixing process over alphabet A with mixing rate function $\beta(\cdot)$. Consider a non-empty finite set $S \subset A^*$ and define:*

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

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

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

then the random variables

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

satisfy:

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

REMARK 9. *In our application we will take $n \geq 3$, $\xi = 1/\log n$, $\delta \geq n^{-\alpha}$ and $|S| \leq n$. In this case the condition on n in the Theorem is satisfied whenever:*

$$n \geq C(1 + \alpha) \left(\frac{(h_S + 1) \log^4 n}{\pi_S} \right) \vee \left(\frac{\log^3 n}{\pi_S} \beta^{-1} \left(\frac{\xi \pi_S \delta}{24} \right) \right)$$

where $C > 0$ is universal.

PROOF OF THEOREM 17. Consider a number $b \in \mathbb{N} \setminus \{0\}$. Given $r \in \mathbb{N}$, a sequence $B = (B_1, \dots, B_r) \in [n]$ of subsets of $[n]$ is said to consist of b -separated blocks if each B_i is an interval in $[n]$ and $\min B_{i+1} \geq \max B_i + b$ for $1 \leq i \leq r-1$. We say that such a sequence is t -regular if $t = |B_1| = |B_2| = \dots = |B_{r-1}| \geq |B_r|$.

LEMMA 6. *Under the assumptions of the Theorem, let (B_1, \dots, B_r) be a sequence of b -separated t -regular blocks where $t \geq 2|w|$. Define for each $w \in S$ the number of occurrences of w that are contained in one of the blocks B_i :*

$$N(w) \equiv |\{j \in [n] : \exists i \in [r+1], j, j + |w| - 1 \in B_i \text{ and } X_j^{j+|w|-1} = w\}|$$

and let n_w denote the number of places where w may occur:

$$n_w \equiv \sum_{i=1}^r (|B_i| - |w| + 1)_+ = (t - |w| + 1)(r - 1) + (|B_r| - |w| + 1)_+.$$

Given $\lambda > 0$, let $E(\lambda)$ denote the event:

$$E(\lambda) \equiv \left\{ \forall w \in S, \left| \frac{N(w)}{n_w} - \pi(w) \right| \leq \lambda \pi(w) \right\}$$

Then

$$\mathbb{P}(E(\lambda)) \geq 1 - 2|S| \exp\left(-\frac{\lambda^2 \pi_S t (r-1)}{16b(1 + \frac{\lambda}{6})}\right) - \frac{2\beta(b)}{\lambda \pi_S}.$$

PROOF OF LEMMA 6. Let $\tilde{X}_{B_1}, \dots, \tilde{X}_{B_{r+1}}$ be a sequence of independent random variables where each \tilde{X}_{B_i} has the same distribution as X_{B_i} . Define $\tilde{N}(w)$ and $\tilde{E}(\cdot)$ in analogy with $N(w)$ and $E(\cdot)$ (respectively).

Our first major goal in the proof is to show:

$$\text{CLAIM 5. } \mathbb{P}(E(\lambda)) \geq \mathbb{P}\left(\tilde{E}(\lambda/2)\right) - \frac{2\beta(b)}{\pi_S \lambda}.$$

To prove the claim we first construct a coupling of $\tilde{X}_{B_1}, \dots, \tilde{X}_{B_r}$ to the process $X_{-\infty}^{+\infty}$. Set $\tilde{X}_{B_1} = X_{B_1}$. Assuming that we have defined \tilde{X}_{B_i} for $1 \leq i < j$, we sample $(X_{B_j}, \tilde{X}_{B_j})$ from a coupling achieving total variation distance. That is to say,

$$\mathbb{P}\left(X_{B_j} \neq \tilde{X}_{B_j} \mid X_{B_i}, i < j\right) = \sup_{E \subset A^{B_j}} \left| \mathbb{P}\left(X_{B_j} \in E \mid X_{B_i}, i < j\right) - \mathbb{P}\left(X_{B_j} \in E\right) \right|.$$

The b separation condition implies that X_{B_j} is b steps ahead into the future from $X_{B_i}, i < j$. Therefore, the β -mixing condition implies:

$$\mathbb{P}\left(X_{B_j} \neq \tilde{X}_{B_j}\right) = \mathbb{E} \left[\sup_{E \subset A^{B_j}} \left| \mathbb{P}\left(X_{B_j} \in E \mid X_{B_i}, i < j\right) - \mathbb{P}\left(X_{B_j} \in E\right) \right| \right] \leq \beta(b).$$

Now observe that in order for $E(\lambda)$ to hold it suffices that $\tilde{E}(\lambda/2)$ holds and that

$$\forall w \in S, \frac{|N(w) - \tilde{N}(w)|}{n_w} \leq \frac{\lambda \pi_S}{2}.$$

Therefore we will be done once we show that:

$$(G.1) \quad \mathbb{P}\left(\forall w \in S, \frac{|N(w) - \tilde{N}(w)|}{n_w} \leq \frac{\lambda \pi_S}{2}\right) \geq 1 - \frac{2\beta(b)}{\lambda \pi_S}.$$

To do this, notice that for any w :

$$|N(w) - \tilde{N}(w)| \leq \sum_{i=1}^r (|B_i| - |w| + 1)_+ \chi_{\{X_{B_i} \neq \tilde{X}_{B_i}\}}.$$

This is because each block B_i may contain at most $|B_i| - |w| + 1$ occurrences of w .

The first $r - 1$ blocks have the same size $|B_i| = t$, whereas the last one cannot be larger, hence $n_w \geq (r - 1)(t - |w| + 1)_+$. Moreover, $X_{B_1} = \tilde{X}_{B_1}$ always. We deduce:

$$|N(w) - \tilde{N}(w)| \leq (t - |w| + 1)_+ \sum_{i=2}^r \chi_{\{X_{B_i} \neq \tilde{X}_{B_i}\}} \leq \frac{n_w}{r - 1} \sum_{i=2}^r \chi_{\{X_{B_i} \neq \tilde{X}_{B_i}\}},$$

and

$$\begin{aligned} \mathbb{E} \left[\max_{w \in S} \frac{|N(w) - \tilde{N}(w)|}{n_w} \right] &\leq \mathbb{E} \left[\max_{w \in S} \frac{|N(w) - \tilde{N}(w)|}{n_w} \right] \\ &\leq \frac{\sum_{i=2}^r \mathbb{P}\left(X_{B_i} \neq \tilde{X}_{B_i}\right)}{r - 1} \\ &\leq \beta(b). \end{aligned}$$

We deduce from Markov's inequality that:

$$\mathbb{P}\left(\max_{w \in S} \frac{|N(w) - \tilde{N}(w)|}{n_w} > \frac{\lambda \pi_S}{2}\right) \leq \frac{2\beta(b)}{\lambda \pi_S},$$

and this is precisely (G.1), which finishes the end of the proof of the Claim.

We must now bound $\mathbb{P}(\tilde{E}(\lambda))$. By the union bound, we have:

$$(G.2) \quad \mathbb{P}(\tilde{E}(\lambda_w)) \geq 1 - \sum_{w \in S} \mathbb{P}\left(\left|\frac{\tilde{N}(w)}{n_w} - \pi(w)\right| > \frac{\lambda \pi(w)}{2}\right).$$

Fix a $w \in S$. Let $\tilde{N}_i(w)$ denote the number of occurrences of w in B_i . We will apply Bennett's inequality to the sum of these random variables. To this end we note that:

1. $\sum_{i=1}^r \tilde{N}_i(w) = \tilde{N}(w)$.
2. *The $\tilde{N}_i(w)$ are independent.* This is so because the \tilde{X}_{B_i} are independent.
3. $\tilde{N}_i(w) \leq (|B_i| - |w| + 1)_+ \leq b$ for all i because b is an upper bound on $|B_i|$.
4. $\sum_i \mathbb{E}[\tilde{N}_i(w)] = \pi(w)n_w$.
5. $\sum_i \mathbb{V}(\tilde{N}_i(w)) \leq \pi(w)n_w b$. This is so because each $\tilde{N}_i(w)$ is a sum of $(|B_i(w)| - |w| + 1)_+ \leq b$ indicators with variance $\pi(w)(1 - \pi(w))$ and the variance of a sum of $\leq b$ terms is at most b times the sum of the variances (by Cauchy Schwarz).

Therefore,

$$\mathbb{P}\left(|\tilde{N}(w) - \pi(w)n_w| \geq \frac{\lambda \pi(w)n_w}{2}\right) \leq 2 \exp\left(-\frac{\lambda^2 \pi(w)n_w}{8b(1 + \frac{\lambda}{6})}\right).$$

Since $t \geq 2h$,

$$\forall w \in S, n_w = \sum_{i=1}^r (|B_i(w)| - |w| + 1)_+ \geq (r-1)(t - |w| + 1) \geq \frac{(r-1)t}{2},$$

and the result follows from plugging the probability inequality into (G.2) and applying the Claim. \square

From now on we set:

$$(G.3) \quad b \equiv \left\lceil \frac{10h_S}{\xi} \right\rceil \vee \beta^{-1} \left(\frac{\xi \pi_S \delta}{24} \right),$$

$$(G.4) \quad r \equiv \left\lceil \frac{n}{2b} \right\rceil.$$

We now construct three sets of b -separated blocks in $[m]$. The first one is:

1. $B^{(1)} = (B_1^{(1)}, \dots, B_r^{(1)})$ consists of intervals of the form

$$B_i^{(1)} \equiv \{b(2i-2) + s : 1 \leq s \leq b\} \cap [m].$$

These are the intervals of length b whose right endpoints are even multiples of b . These are b -separated b -regular blocks.

2. $B^{(2)} = (B_1^{(2)}, \dots, B_r^{(2)})$ consists of intervals of the form

$$B_i^{(2)} \equiv \{b(2i-1) + s : 1 \leq s \leq b\} \cap [m].$$

These are the intervals whose right endpoints are odd multiples of b .

In this case we set $B_i^{(2)}(w) = B_i^{(2)}$ for each $1 \leq i \leq r$ and $w \in S$. This also results in b -separated b -regular blocks.

3. $B^{(3)} = (B_i^{(3)}, \dots, B_{2r-1}^{(3)})$ consists of intervals

$$B_i^{(3)} = \{bi - h_S + 2, bi - h_S + 3, \dots, bi + h_S\} \cap [m].$$

This results in b -separated $2h_S$ -regular blocks, as one can check. (Here one must use $b \geq 2h_S$, which follows from (G.3) and the assumption $\xi < 1/2$.)

For each $k \in \{1, 2, 3\}$, let $N^{(k)}(w)$ count the number of occurrences of w that are contained in a block of the form $B_i^{(k)}$ and let $n_w^{(k)} = \sum_i (|B_i^{(k)}| - |w| + 1)_+$. We will need two propositions.

PROPOSITION 5. *For any $w \in S$,*

$$N^{(1)}(w) + N^{(2)}(w) \leq N_m(w) \leq N^{(1)}(w) + N^{(2)}(w) + N^{(3)}(w).$$

PROOF. The LHS counts the number of occurrences of w contained in intervals of the form $\{bi + 1, bi + 2, \dots, b(i + 1)\}$. Since $|w| \leq h_S$, $N^{(3)}(w)$ is an upper bound on the number of occurrences of w that are not entirely contained in one of those intervals. \square

PROPOSITION 6. For any $w \in S$, $n_w^{(1)}, n_w^{(2)} \geq \frac{(1-\xi/3)n}{2}$, $n_w^{(1)} + n_w^{(2)} \leq n$ and $n_w^{(3)} \leq \frac{3nh_S}{b}$.

PROOF. Since $B^{(1)}$ is b -regular and $r \geq n/2b$ (by G.4),

$$n_w^{(1)} \geq (r-1)(b-|w|) \geq \frac{n}{2} - \frac{n|w|}{2b} - b = \frac{n}{2} \left(1 - \frac{h_S}{b} - \frac{b}{n}\right).$$

By (G.3) $b \geq 6h_S/\xi$ and $n \geq 6b/\xi$, so $n_1^{(w)} \geq (1-\xi/3)n/2$ as desired. The same argument works for $n_w^{(2)}$. For $n_w^{(3)}$ we start from $2r-1 \leq 2(n/2b+1)-1 = n/2b+1$. Since each block contains at most $2h_S$ points,

$$n_w^{(3)} \leq \frac{nh_S}{b} + 2h_S.$$

The rest follows from $b/n \leq \xi < 1$. \square

Consider the events:

$$(G.5) \quad G^{(1)} \equiv \left\{ \forall w \in S, \left| \frac{N^{(1)}(w)}{n_w^{(1)}} - \pi(w) \right| < \frac{\xi \pi(w)}{3} \right\};$$

$$(G.6) \quad G^{(2)} \equiv \left\{ \forall w \in S, \left| \frac{N^{(2)}(w)}{n_w^{(2)}} - \pi(w) \right| < \frac{\xi \pi(w)}{3} \right\};$$

$$(G.7) \quad G^{(3)} \equiv \left\{ \forall w \in S, \frac{N^{(3)}(w)}{n} \leq \frac{2\xi \pi(w)}{3} \right\};$$

$$(G.8) \quad G = G^{(1)} \cap G^{(2)} \cap G^{(3)}.$$

CLAIM 6. $G \subset \{\forall w \in S : |N_m(w) - \pi(w)n| \leq \xi \pi(w)n\}$.

PROOF. Assume G holds. Then for any $w \in S$:

$$\begin{aligned} N_m(w) &\geq N^{(1)}(w) + N^{(2)}(w) \\ (G \text{ occurs}) &\geq \pi(w) \left(1 - \frac{\xi}{3}\right) (n_w^{(1)} + n_w^{(2)}) \\ (\text{Proposition 6}) &\geq \pi(w) \left(1 - \frac{\xi}{3}\right)^2 n \\ &\geq (1-\xi) \pi(w) n. \end{aligned}$$

On the other hand,

$$\begin{aligned} N_m(w) &\leq N^{(1)}(w) + N^{(2)}(w) + N^{(3)}(w) \\ (G \text{ occurs}) &\leq \pi(w) \left(1 + \frac{\xi}{3}\right) n + \frac{2\xi \pi(w)}{4} n \\ &\leq (1+\xi) \pi(w) n. \end{aligned}$$

□

The claim implies that:

$$\mathbb{P}(G) \leq \mathbb{P}(\forall w \in S : |N_m(w) - \pi(w)n| \leq \xi\pi(w)n)$$

and we proceed to bound $\mathbb{P}(G)$.

We first apply Lemma 6 to each $G^{(k)}$. For $k = 1, 2$ we may take $t = b$, $\lambda = \xi/3$ and note that $r \geq n/2b$, $\beta(b) \leq \lambda\delta\pi_S/8$ (cf. G.4, G.3) to deduce:

$$(G.9) \quad \mathbb{P}\left((G^{(1)})^c\right) + \mathbb{P}\left((G^{(2)})^c\right) \leq 4|S| \exp\left(-\frac{\xi^2 \pi_S (r-1)}{144 \left(1 + \frac{\xi}{18}\right)}\right) + \frac{\delta}{4}.$$

For $G^{(3)}$ we take $t = 2h_S$ and

$$\lambda = \frac{2\xi n}{3 \max_{w \in S} n_w} - 1 \geq \frac{2\xi b}{9h_S} - 1 \geq \frac{\xi b}{9h_S}$$

since $b \geq 10h_S/\xi$ by (G.3). Notice that in this case $\lambda > 1$, hence

$$\lambda^2/(1 + \lambda/6) \geq 3\lambda/4 \geq \frac{\xi b}{12h_S}.$$

Moreover, $\beta(b)/\pi_S \lambda \leq \delta/6$. Hence, we deduce:

$$(G.10) \quad \begin{aligned} \mathbb{P}\left((G^{(3)})^c\right) &\leq 2|S| \exp\left(-\frac{\pi_S 2h_S(r-1)}{16b} \frac{\lambda^2}{1+\lambda/6}\right) + \frac{\delta}{6} \\ &\leq 2|S| \exp\left(\frac{\xi \pi_S (r-1)}{192}\right) + \frac{\delta}{6}. \end{aligned}$$

Now compare the exponential terms in the two equations. Since $\xi \leq 1/2$,

$$\frac{150}{\xi^2} \geq \frac{144 \left(1 + \frac{\xi}{18}\right)}{\xi^2} \geq \frac{288}{\xi} \geq \frac{192}{\xi},$$

hence the exponential term in (G.10) is larger than the exponential term in (G.9). We conclude that:

$$\mathbb{P}(G) \leq 1 - \sum_{k=1}^3 \mathbb{P}\left((G^{(k)})^c\right) \geq 1 - \frac{\delta}{2} - 6|S| \exp\left(-\frac{r-1}{\frac{300}{\xi^2 \pi_S}}\right).$$

To finish the proof, we recall that $r \geq n/2b$ (cf. G.4) and notice that our assumptions imply:

$$\frac{n}{2b} \geq 1 + \frac{300}{\xi^2 \pi_S} \ln\left(\frac{12|S|}{\delta}\right).$$

Plugging this back into the previous inequality gives $\mathbb{P}(G) \geq 1 - \delta$ and finishes the proof. □

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.

REFERENCES

- [1] V. Aguirregabiria and P. Mira. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156:38–67, 2010.
- [2] M. Arellano and B. H. Honoré. Panel data models: Some recent developments. *Handbook of Econometrics*, 5:3229–3296, 2001.
- [3] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.
- [4] M. Browning and J. M. Carro. Heterogeneity in dynamic discrete choice models. *Econometrics Journal*, 13(1):1–39, 2010.
- [5] M. Browning and J. M. Carro. Dynamic binary outcome models with maximal heterogeneity. 2011.
- [6] P. Bühlmann. Efficient and adaptive post-model-selection estimators. *Journal of Statistical Planning and Inference*, 79(1):1–9, 1999.
- [7] P. Bühlmann. Model selection for variable length markov chains and tuning the context algorithm. *Ann. Inst. Statist. Math.*, 52(2):287–315, 2000.
- [8] P. Bühlmann and A. J. Wyner. Variable length markov chains. *Annals of Statistics*, 27(2):480–513, 1999.
- [9] V. Chernozhukov, I. Fernandez-Val, J. Hahn, and W. Newey. Identification and estimation of marginal effects in nonlinear panel models. *arXiv:0904.1990*, 2009.
- [10] F. Comets, R. Fernández, and P. A. Ferrari. Processes with long memory: Regenerative construction and perfect simulation. *The Annals of Applied Probability*, 12(3):921–943, 2002.
- [11] I. Csiszár. Large-scale typicality of markov sample paths and consistency of mdl order estimators. *IEEE Transactions on Information Theory*, 48(6):1616–1628, 2002.
- [12] I. Csiszár and P. Shields. The consistency of the bic markov order estimator. *Ann. Statist.*, 28:16011619, 2000.
- [13] I. Csiszár and P. C. Shields. Redundancy rates for renewal and other processes. *IEEE Transactions on Information Theory*, 42(6):2065–2072, 1996.
- [14] I. Csiszár and Z. Talata. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Trans. Inform. Theory*, 52:1007–1016, 2006.
- [15] F. Ferrari and A. J. Wyner. Estimation of general stationary processes by variable length markov chains. *Scandinavian Journal of Statistics*, 30:459–480, 2003.
- [16] A. Galves, C. Galves, N. Garcia, and F. Leonardi. Context tree selection and linguistic rhythm retrieval from written texts. *ArXiv*, (0902.3619v2), 2009.
- [17] A. Garivier. Redundancy of the context-tree weighting method on renewal and markov renewal processes. *IEEE Transactions on Information Theory*, 52:5579–5586, 2006.
- [18] A. Garivier and F. Leonardi. Context tree selection: A unifying view. *arXiv:1011.2424v2*, 2010.
- [19] S. Kalikow. Random Markov processes and uniform martingales. *Israel Journal of Mathematics*, 71(1):33–54, 1990.
- [20] M. Ledoux and M. Talagrand. *Probability in Banach Spaces (Isoperimetry and processes)*. Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer-Verlag, 1991.

- [21] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *arXiv:0903.1468v1 [stat.ML]*, 2010.
- [22] M. L. Putterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.
- [23] J. Rissanen. A universal data compression system. *IEEE Tran. Inform. Theory*, 29:656–664, 1983.
- [24] S. M. Ross. *Introduction to stochastic dynamic programming*. Academic Press, 1983.
- [25] S. M. Ross. *Dynamic programming: deterministic and stochastic models*. Prentice-Hall, Inc., 1987.
- [26] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated markov models. *Nucleic Acids Research*, 26(2):544–548, 1998.
- [27] Z. Talata and T. Duncan. Unrestricted bic context tree estimation for not necessarily finite memory processes. *ISIT*, pages 724–728, 2009.
- [28] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:49–67, 2006.

100 FUQUA DRIVE
DURHAM, NC 27708
E-MAIL: abn5@duke.edu

ESTRADA DONA CASTORINA 110
RIO DE JANEIRO, RJ
E-MAIL: rinfo@impa.br