Convergence of Binarized Context-tree Weighting for Estimating Distributions of Stationary Sources

Badri N. Vellambi, Marcus Hutter
(Presented by **Parastoo Sadeghi**)

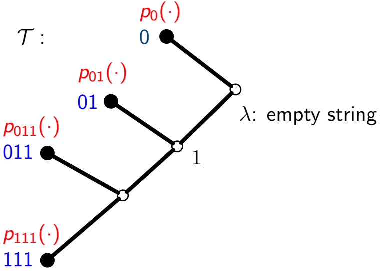
Australian National University

June 18, 2018

International Symposium on Information Theory
Vail, CO

Tree Sources

- k-order Markov Source: $p_n(X_1, \ldots, X_n) = \pi(X_1, \ldots, X_k) \prod_{j=k+1}^{n} p(X_j | X_{j-1}, \ldots X_{j-k})$
  - Defined uniquely by π and k-**order** conditional distribution p.
- Tree (or Variable-order Markov) Sources:
  - The # of RVs in the conditioning of the product term varies with the realization.
  - Defined by: (a) a **complete context** tree [i.e., leaves form a sufix-free code and satisfy Kraft's inequality]; and (b) appropriate variable-order conditional distributions.



$$p_0(\cdot) := p(X_i = \cdot | X_{i-1} = 0)$$
$$p_{01}(\cdot) := p(X_i = \cdot | X_{i-1}X_{i-2} = 01)$$
$$p_{011}(\cdot) := p(X_i = \cdot | X_{i-1}X_{i-2}X_{i-1} = 011)$$
$$p_{111}(\cdot) := p(X_i = \cdot | X_{i-1}X_{i-2}X_{i-1} = 111)$$

$$p_{\mathcal{T}}(0111111) = \pi(0) p_0(1) p_{01}(1) p_{011}(1) p_{111}^3(1)$$
$$p_{\mathcal{T}}(0000000) = \pi(0) p_0^6(0)$$

(Binary) Context-tree Weighting (CTW) Method

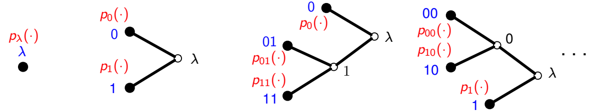
- CTW estimate is a Bayesian mixture of tree source estimates.

$$p_{CTW}(x_{1:n}) = \sum_{\mathcal{T}} \omega_{\mathcal{T}} p_{\mathcal{T}}(x_{1:n}),$$

where $\omega_{\mathcal{T}} > 0$ for every context tree, and for context $\boldsymbol{a} \in \mathcal{T}$,

$$p_{\boldsymbol{a}}(0) = 1 - p_{\boldsymbol{a}}(1) := \frac{\#\boldsymbol{a}0 + \frac{1}{2}}{\#\boldsymbol{a} + 1} \qquad \text{(add-half or Laplace estimator)}$$

$$\#\boldsymbol{s} = \text{the number of times the string } \boldsymbol{s} \text{ appears in } x_{1:n}.$$



IID source 1-order Markov Source

[Willems et al.-T-IT95]

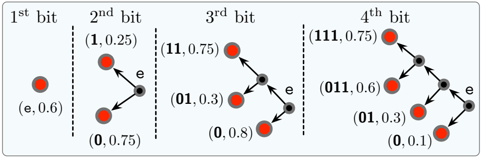
- CTW yields an optimal, consistent, adaptive, strongly-sequential estimate of (stationary) distributions of tree sources.
- Worst-case redundancy bounds: For any tree source $\mathcal{T}$ with C leaves/contexts,

$$\max_{x_{1:n}} \rho(x_{1:n}) := \max_{x_{1:n}} \log_2 \frac{p_{\mathcal{T}}(x_{1:n})}{p_{CTW}(x_{1:n})} \leq \mathrm{C}\left(\tfrac{1}{2}\log_2 \tfrac{n}{\mathrm{C}} + 1\right) + (2\mathrm{C} - 1) + 2.$$

CTW Extensions for Tree Sources with Non-binary Alphabets

- Straightforward generalization of CTW to non-binary (tree) sources is sub-optimal [Tjalkens et al-ISIT'93]
- Extensions of CTW for non-binary tree sources using hierarchical decomposition/binarization of the alphabet was proposed [Tjalkens et al-ISIT'94, Tjalkens et al-DCC'97]

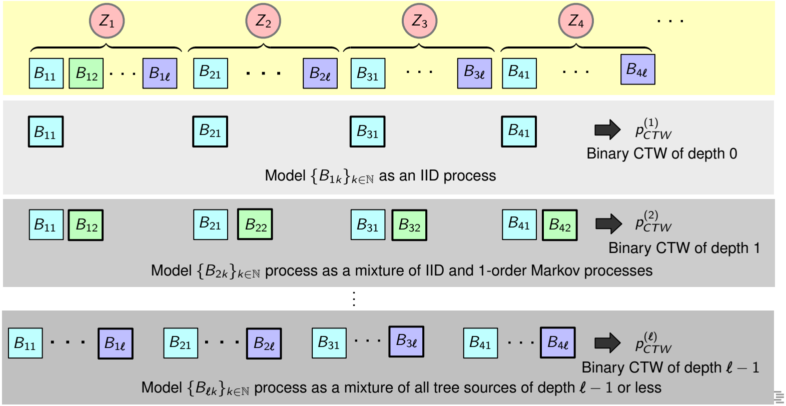
0	$9/250$	4	$9/2000$	8	$1/125$	12	$9/400$
1	$81/250$	5	$81/2000$	9	$9/125$	13	$81/400$
2	$27/1000$	6	$63/1000$	10	$3/500$	14	$9/160$
3	$63/1000$	7	$21/500$	11	$7/500$	15	$3/160$



- binarization allows the possibility of exploiting any (tree) structure of the correlation between component bits, which a naïve non-binary CTW cannot exploit.

Binarized CTW

- Recently, a binarized CTW approach was used to estimate the underlying stationary distribution of a hidden Markov model process [Veness et al.-AAAI'15]
- This approach translates the problem of policy evaluation and on-policy control in reinforcement learning to estimation (of stationary distribution).
- The binarized CTW translates estimation of the stationary distribution over 2^ℓ symbols to those of ℓ binary sources as follows.



Binarized CTW

$$\hat{P}_{\overline{CTW}}(z_{1:n}) := \prod_{i=1}^{\ell} p_{CTW}^{(i)}(z_{1:n})$$

- $\hat{P}_{\overline{CTW}}$ is a product of ℓ component binary CTW estimates.
- $p_{CTW}^{(i)}$ is a binary CTW estimate that depends only on the first i component binary processes, i.e., $\{B_{kj} : 1 \leq k \leq i, 1 \leq j \leq n\}$.
- Simulations in [veness et al.-AAAI'15] reveal that $\hat{p}_{\overline{CTW}}$ has:
  - excellent convergence rate in estimating the stationary distribution of the underlying process;
  - the ability to handle much larger alphabets than the frequency estimator.

In this work, we...

- show that the worst-case L_1-prediction error between the binarized CTW and frequency (ML) estimates for the stationary distribution of a stationary ergodic source over $\{0;1\}^{\ell}$ for some $\ell > 1$ is $\Theta\left(\sqrt{2^{\ell} \frac{\log n}{n}}\right)$.
- (consequently,) establish the consistency of the binarized CTW estimator

Main Results

Let $\hat{P}_{\overline{\text{CTW}}}(\boldsymbol{c}\,; z_{1:n})$ denote the binarized CTW estimate of the distribution of a random process Z after observing n symbols of the process, i.e.,

$$\hat{P}_{\overline{\text{CTW}}}(\boldsymbol{c}; z_{1:n}) := \frac{\hat{p}_{\overline{\text{CTW}}}(z_{1:n}\boldsymbol{c})}{\hat{p}_{\overline{\text{CTW}}}(z_{1:n})}, \quad \boldsymbol{c} \in \mathcal{Z}, z_{1:n} \in \mathcal{Z}^n$$

Theorem (Lower Bound)

Let $\mathcal{Z} = \{0,1\}^\ell$ for $\ell \geq 2$ denote the alphabet of a given random process. Then, for $\epsilon > 0$, there exist $n \in \mathbb{N}$ and $z_{1:n} \in \mathcal{Z}^n$ such that

$$\Delta_\ell := \sum_{\boldsymbol{c}\in\{0,1\}^\ell} \left| \hat{P}_{\overline{CTW}}(Z = \boldsymbol{c}\,; z_{1:n}) - \frac{\#_\ell \boldsymbol{c}}{n} \right| \geq \sqrt{\frac{2^{\ell-2}(1-\epsilon)\log n}{n}}.$$

Theorem (Upper Bound)

Let $\mathcal{Z} = \{0,1\}^\ell$ for $\ell \in \mathbb{N}$ denote the alphabet of a given random process. Then,

$$\Delta_\ell := \max_{z_{1:n}\in\mathcal{Z}^n} \sum_{\boldsymbol{c}\in\{0,1\}^\ell} \left| \hat{P}_{\overline{CTW}}(Z = \boldsymbol{c}\,; z_{1:n}) - \frac{\#_\ell \boldsymbol{c}}{n} \right| \leq \begin{cases} \frac{1}{n} & \ell = 1 \\ \Delta_{\ell-1} + \frac{2^{\ell-1}}{2n} + \sqrt{\frac{2^{\ell-2}}{n} \log\left(\frac{2\pi e^5 n}{2^{\ell-1}}\right)} & \ell > 1 \end{cases}.$$

Outline of Lower Bound

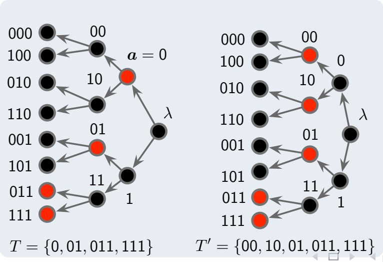
- Identify explicitly a sequence $z_{1:n}$ for which the lower bound holds.
- Let $n = 2^{\ell} m$ and $\sigma > 0$.
- Let $z_{1:n}$ be a sequence such that the number of occurrence of $\boldsymbol{a} \in \{0,1\}^{\ell}$ is

$$\#\boldsymbol{a} = \begin{cases} m - \lfloor \sigma\sqrt{m \log m} \rfloor & \boldsymbol{a} \text{ is of even weight} \\ m + \lfloor \sigma\sqrt{m \log m} \rfloor & \boldsymbol{a} \text{ is of odd weight} \end{cases} .$$

- The frequency of symbols is nearly equiprobable, but deviates from the equiprobable distribution by a factor that is $\Theta\left(\sqrt{\frac{\log n}{n}}\right)$.
- The proof proceeds by computing the two estimates to show explicitly that the lower bound holds for this choice of frequencies.

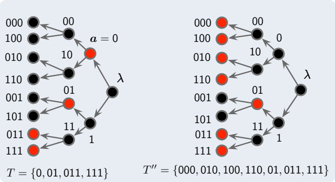
Outline of Upper Bound

- Proof follows by induction.
- Since the binarized CTW estimate is a product of ℓ binary CTW estimates, one needs to identify the dominant tree source in the Bayesian mixture corresponding to each of the binary CTW estimates.
- Consider the binary CTW estimate for the k^{th} bit.
- To identify the dominant tree in this estimate, we need to compare the contribution of each context tree in each Bayesian mixture; this is done in three steps.
- Step 1: Compare the contributions of two trees T, T' such that $T' = (T \setminus \{\boldsymbol{a}\}) \cup \{0\boldsymbol{a}, 1\boldsymbol{a}\}$ for some $\boldsymbol{a} \in T$.



Outline of Upper Bound

- Step 2: By repeated use of Step 1, compare the contributions of two trees T, T'' such that $T'' = (T \setminus \{\boldsymbol{a}\}) \cup \{\text{all leaves with suffix } \boldsymbol{a}\}$



- Step 3; By repeated use of Step 2, one can compare the contributions of a tree T and the context-tree T^*_{k-1} corresponding to the $(k-1)$-order Markov source to show that:

$$\frac{p_T^{(k)}(z_{1:n})}{p_{T^*_{k-1}}^{(k)}(z_{1:n})} \leq \frac{(2\pi)^{\frac{2^{k-1}-|T|}{2}} \left(\prod_{\boldsymbol{a} \in T : \#_{k-1}\boldsymbol{a} > 0} \frac{\lambda_{\boldsymbol{a}}}{\sqrt{\#_{k-1}\boldsymbol{a}}} \right) \prod_{\boldsymbol{c} \in \{0,1\}^{k-1} : \#_{k-1}\boldsymbol{c} > 0} \sqrt{\#_{\ell-1}\boldsymbol{c}}}{\exp\left\{ 4 \sum_{\boldsymbol{a} \in T} \sum_{\boldsymbol{b} \in \{0,1\}^{m_{\boldsymbol{a}}} : \#_{k-1}\boldsymbol{ba} > 0} \#_{k-1}\boldsymbol{ba} \left(\frac{\#_\ell \boldsymbol{ba}0}{\#_{k-1}\boldsymbol{ba}} - \frac{\#_k \boldsymbol{a}0}{\#_{k-1}\boldsymbol{a}} \right)^2 \right\}},$$

where $m_{\boldsymbol{a}} := k - 1 - \|\boldsymbol{a}\|$, and $|\log \lambda_{\boldsymbol{a}}| \leq \frac{2^{m_{\boldsymbol{a}}}-1}{2}$.

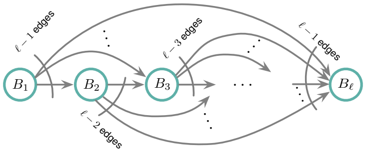
Outline of Upper Bound

$$\frac{p_T^{(k)}(z_{1:n})}{p_{T_{k-1}^*}^{(k)}(z_{1:n})} \leq \frac{(2\pi)^{\frac{2^{k-1}-|T|}{2}} \left(\prod_{\boldsymbol{a} \in T : \#_{k-1}\boldsymbol{a} > 0} \frac{\lambda_{\boldsymbol{a}}}{\sqrt{\#_{k-1}\boldsymbol{a}}} \right) \prod_{\boldsymbol{c} \in \{0,1\}^{k-1} : \#_{k-1}\boldsymbol{c} > 0} \sqrt{\#_{\ell-1}\boldsymbol{c}}}{\exp\left\{ 4 \sum_{\boldsymbol{a} \in T} \sum_{\boldsymbol{b} \in \{0,1\}^{m_{\boldsymbol{a}}} : \#_{k-1}\boldsymbol{ba} > 0} \#_{k-1}\boldsymbol{ba} \left(\frac{\#_{\ell}\boldsymbol{ba}0}{\#_{k-1}\boldsymbol{ba}} - \frac{\#_{k}\boldsymbol{a}0}{\#_{k-1}\boldsymbol{a}} \right)^2 \right\}},$$

- The largest the numerator can grow is polynomially in n
- The largest the denominator can grow is exponentially in n.
- If simpler model T explains data $z_{1:n}$ better than the complicated model T_{k-1}^*, the contribution of T can be larger than that of T_{k-1}^* only by a polynomial factor.
- If complicated model T_{k-1}^* explains data $z_{1:n}$ better than the simpler model T, the contribution of T_{k-1}^* can be larger than that of T by an exponential factor.
- Upon rearranging terms, and relating the required L_1-predictive error between the binarized CTW and ML estimates to the denominator term yields the upper bound.

Additional Structure between Component Bits

- The binarized CTW presented so far assumes no known structure between component bits, i.e., the k^{th} bit is assumed to depend on all previous $k-1$ component bits.



- If it is known that the component bits satisfy some structure (given by a Bayesian Network $\mathcal{B}$), we can incorporate accordingly to derive a suitable binarized CTW estimate $\hat{P}^{\mathcal{B}}_{\overline{CTW}}$; Corresponding to our previous results, we can show that:

Theorem

Given Bayesian network $\mathcal{B}$ consisting of k binary random variables, and $\mathcal{Z} = \{0,1\}^\ell$,

$$\max_{z_{1:n}\in\mathcal{Z}^n} \left\| \hat{P}^{\mathcal{B}}_{\overline{CTW}}(\cdot\,; z_{1:n}) - \hat{P}_{\mathrm{ML},\mathcal{B}}(\cdot\,; z_{1:n}) \right\| = \Theta\left(\sqrt{\frac{\log n}{n}}\right),$$

where $\hat{P}_{\mathrm{ML},\mathcal{B}}(\cdot\,; z_{1:n}) := \underset{P \text{ satisfies } \mathcal{B}}{\operatorname{argmax}} P(z_{1:n})$,