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
Some Preliminary Concepts

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 $\pi$ and *k-order* conditional distribution $p$.

- **Tree (or Variable-order Markov) Sources:**
  
  - The number 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 suffix-free code and satisfy Kraft’s inequality]; and (b) appropriate variable-order conditional distributions.

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

\[
p_T(0111111) = \pi(0)p_0(1)p_{01}(1)p_{011}(1)p_{111}^3(1) \\
p_T(0000000) = \pi(0)p_0^6(0)
\]
Some Preliminary Concepts

(Binary) Context-tree Weighting (CTW) Method

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

\[ p_{CTW}(x_1:n) = \sum_{T} \omega_T p_T(x_1:n), \]

where \( \omega_T > 0 \) for every context tree, and for context \( a \in \mathcal{T} \),

\[ p_a(0) = 1 - p_a(1) := \frac{\#a0 + \frac{1}{2}}{\#a + 1} \quad \text{(add-half or Laplace estimator)} \]

\( \#s \) = the number of times the string \( s \) appears in \( x_1:n \).

- 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,

\[
\max_{x_1:n} \rho(x_1:n) := \max_{x_1:n} \log_2 \frac{p_{\mathcal{T}}(x_1:n)}{p_{CTW}(x_1:n)} \leq C \left( \frac{1}{2} \log_2 \frac{n}{C} + 1 \right) + (2C - 1) + 2.
\]

[Willems et al.-T-IT95]
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]

<table>
<thead>
<tr>
<th>0</th>
<th>9/250</th>
<th>4</th>
<th>9/2000</th>
<th>8</th>
<th>1/125</th>
<th>12</th>
<th>9/400</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81/250</td>
<td>5</td>
<td>81/2000</td>
<td>9</td>
<td>9/125</td>
<td>13</td>
<td>81/400</td>
</tr>
<tr>
<td>2</td>
<td>27/1000</td>
<td>6</td>
<td>63/1000</td>
<td>10</td>
<td>3/500</td>
<td>14</td>
<td>9/160</td>
</tr>
<tr>
<td>3</td>
<td>63/1000</td>
<td>7</td>
<td>21/500</td>
<td>11</td>
<td>7/500</td>
<td>15</td>
<td>3/160</td>
</tr>
</tbody>
</table>

- 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 $2^\ell$ symbols to those of $\ell$ binary sources as follows.

Model $\{B_{1k}\}_{k \in \mathbb{N}}$ as an IID process

Model $\{B_{2k}\}_{k \in \mathbb{N}}$ process as a mixture of IID and 1-order Markov processes

Model $\{B_{\ell k}\}_{k \in \mathbb{N}}$ process as a mixture of all tree sources of depth $\ell - 1$ or less
Some Preliminary Concepts

Binarized CTW

\[ \hat{P}_{\text{CTW}}(z_1:n) := \prod_{i=1}^{\ell} p_{\text{CTW}}^{(i)}(z_1:n) \]

- \( \hat{P}_{\text{CTW}} \) is a product of \( \ell \) component binary CTW estimates.
- \( p_{\text{CTW}}^{(i)} \) 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 \( \hat{p}_{\text{CTW}} \) 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 \( \Theta \left( \sqrt{2\ell \log n \over n} \right) \).
- (consequently,) establish the consistency of the binarized CTW estimator
Main Results

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

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

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

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

Theorem (Upper Bound)

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

$$\Delta_{\ell} := \max_{z_{1:n} \in \mathcal{Z}^n} \sum_{c \in \{0, 1\}^\ell} \left| \hat{P}_{\text{CTW}}(Z = c ; z_{1:n}) - \#c_{n} \right|$$

$$\leq \begin{cases} \frac{1}{n} & \ell = 1 \\ \Delta_{\ell-1} + \frac{2^{\ell-1}}{2n} + \sqrt{\frac{2\ell - 2}{n} \log \left( \frac{2\pi e^5 n}{2^{\ell-1}} \right)} & \ell > 1 \end{cases}.$$
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 $a \in \{0, 1\}^\ell$ is
  \[
  \#a = \begin{cases} 
  m - \lfloor \sigma \sqrt{m \log m} \rfloor & a \text{ is of even weight} \\
  m + \lfloor \sigma \sqrt{m \log m} \rfloor & a \text{ is of odd weight} 
  \end{cases}.
  \]
- The frequency of symbols is nearly equiprobable, but deviates from the equiprobable distribution by a factor that is $\Theta(\sqrt{\frac{\log n}{n}})$.
- The proof proceeds by computing the two estimates to show explicitly that the lower bound holds for this choice of frequencies.
Main Results

Outline of Upper Bound

Proof follows by induction.

Since the binarized CTW estimate is a product of $\ell$ 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 $T' = (T \setminus \{a\}) \cup \{0a, 1a\}$ for some $a \in T$. 

![Diagram]

$T = \{0, 01, 011, 111\}$

$T' = \{00, 10, 01, 011, 111\}$

$a = 0$
Outline of Upper Bound

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

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:

$$\frac{p_T^{(k)}(z_{1:n})}{p_{T_{k-1}^*}^{(k)}(z_{1:n})} \leq \exp \left\{ 4 \sum_{a \in T} \sum_{b \in \{0,1\}^{m_a} : \#_{k-1} ba > 0} \#_{k-1} ba \left( \frac{\#_{\ell} ba_0}{\#_{k-1} ba} - \frac{\#_k a_0}{\#_k a} \right)^2 \right\},$$

where $m_a := k - 1 - \|a\|$, and $|\log \lambda_a| \leq \frac{2^{m_a-1}}{2}$. 
Outline of Upper Bound

\[
\frac{p_T^{(k)}(z_{1:n})}{p_{T^*_{k-1}}^{(k)}(z_{1:n})} \leq \frac{(2\pi)^{\frac{2^{k-1} - |T|}{2}}}{\left(\prod_{a \in T : \#_{k-1}a > 0} \frac{\lambda a}{\sqrt{\#_{k-1}a}}\right) \prod_{c \in \{0,1\}^{k-1} : \#_{k-1}c > 0} \sqrt{\#_{\ell-1}c} \exp \left\{ 4 \sum_{a \in T} \sum_{b \in \{0,1\}} \sum_{m} \#_{k-1}ba \left( \frac{\#_{\ell}ba0}{\#_{k-1}ba} - \frac{\#_{k}a0}{\#_{k-1}a} \right)^2 \right\}}.
\]

- 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 $\hat{P}_{\mathcal{B} \mid CTW}^B$; Corresponding to our previous results, we can show that:

**Theorem**

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

$$\max_{z_1:n \in \mathcal{Z}_n} \left\| \hat{P}_{\mathcal{B} \mid CTW}^B(\cdot; z_1:n) - \hat{P}_{ML, \mathcal{B}}(\cdot; z_1:n) \right\| = \Theta\left(\sqrt{\frac{\log n}{n}}\right),$$

where $\hat{P}_{ML, \mathcal{B}}(\cdot; z_1:n) := \arg\max_{P \text{ satisfies } \mathcal{B}} P(z_1:n)$.