EE378A Statistical Signal Processing Lecture 9 - 05/02/2017

Lecture 9: Context Tree Weighting

Lecturer: Tsachy Weissman *Scribe: Chuan-Zheng Lee*

In the last lecture we have talked abut the mixture idea for i.i.d. sources and Markov sources. Specifically, we introduced the *Krichevsky–Trofimov probability assignment* for a binary vector $x^n \in \{0,1\}^n$ given by

$$P^{\mathrm{KT}}(x^n) = \int \omega(\theta)\theta^{n_1(x^n)}(1-\theta)^{n_0(x^n)}d\theta = \frac{\Gamma(n_0+\frac{1}{2})\Gamma(n_1+\frac{1}{2})}{(\Gamma(\frac{1}{2}))^2\Gamma(n+1)}, \tag{1}$$

where $\omega(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}$, $n_1 \equiv n_1(x^n)$ is the number of 1's in x^n, $n_0 \equiv n_0(x^n)$ is the number of 0's (so that $n = n_0 + n_1$) and $\Gamma(\cdot)$ is the gamma function. Note that $\omega(\theta)$ is a Dirichlet prior with parameters $(\frac{1}{2}, \frac{1}{2})$, and the integrand is just a Dirichlet distribution with parameters $(n_0 + \frac{1}{2}, n_1 + \frac{1}{2})$. The main result is that, for regular models with d degrees of freedom, the Dirichlet mixing idea achieves the optimal worst-case regret $\frac{d}{2}\log n + o(\log n)$ for individual sequences of length n.

In this lecture, we will extend this idea to general *tree sources.*

1 Tree Source

We first begin with an informal defintion of a tree source.

Definition 1 (Context of a symbol)**.** *A* context *of a symbol x_t is a suffix of the sequence that precedes x_t, that is, $(\cdots, x_{t-2}, x_{t-1})$.*

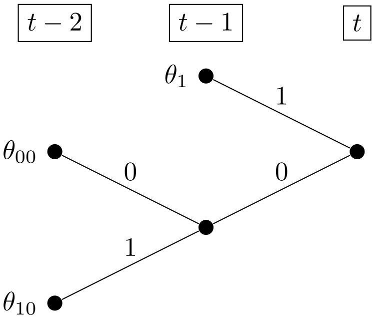
Definition 2 (Tree source)**.** *A* tree source *is a source where a set of suffixes is sufficient to determine the probability distribution of the next item (regardless of the past) in the sequence.*

A tree source is hence a generalization of an order-one Markov source, where only the previous element in the sequence determines the next: now, instead of just one element being sufficient, the *suffix* is sufficient. It is perhaps best elucidated by example.

Example 3. Let $\mathcal{S} = \{00, 10, 1\}$ and $\Theta = \{\theta_s, s \in \mathcal{S}\}$. Then

$$p(X_t = 1|X_{t-2} = 0, X_{t-1} = 0) = \theta_{00}$$
$$p(X_t = 1|X_{t-2} = 1, X_{t-1} = 0) = \theta_{10}$$
$$p(X_t = 1|X_{t-1} = 1) = \theta_1$$

We can represent this structure in a binary tree:



Remark Note that the source in Example 3 is not first-order Markov, but can be cast as a special case of a second-order Markov source.

We denote as $P_{\mathcal{S},\theta}(x^n)$ the associated probability mass function on $x^n \in \mathcal{X}^n$.

2 Known tree, unknown parameters

Recall that with Markov sources, we could think of the pmf of any sequence as the product of conditional pmfs. For the tree sources, we can factorize similarly, but this pertains to suffixes.

Let $P_{\mathcal{S}}^{\mathrm{KT}}$ be the pmf induced by employing the Krichevsky–Trofimov sequence probability assignment (separately) on each subsequence corresponding to every $s \in \mathcal{S}$. We will denote as n_s the number of $x_i, 1 \le i \le n$ with context s, and we will show in Homework 4 that

$$\log \frac{1}{P^{\mathrm{KT}}(x^n)} - \min_{\theta} \log \frac{1}{Q^n_{\mathrm{Bernoulli}(\theta)}(x^n)} \le \frac{1}{2} \log n + 1.$$

Now enumerating over different contexts, the regret on sequence x^n is upper-bounded by

$$\begin{aligned}
\log \frac{1}{P_{\mathcal{S}}^{\mathrm{KT}}(x^n)} - \min_{\theta} \log \frac{1}{P_{\mathcal{S},\theta}(x^n)} &\le \sum_{s \in \mathcal{S}} \left[\frac{1}{2} \log n_s + 1 \right] \\
&= |\mathcal{S}| \sum_{s \in \mathcal{S}} \frac{1}{|\mathcal{S}|} \left[\frac{1}{2} \log n_s + 1 \right] \\
&\overset{\text{(a)}}{\le} |\mathcal{S}| \left[\frac{1}{2} \log \sum_{s \in \mathcal{S}} \frac{1}{|\mathcal{S}|} n_s + 1 \right] \\
&= \frac{|\mathcal{S}|}{2} \log \frac{n}{|\mathcal{S}|} + |\mathcal{S}|
\end{aligned}$$

where (a) follows from Jensen's inequality. Recall from Lecture 8 that Rissannen showed that the worst-case regret could at best be $\frac{d}{2} \log n + o(\log n)$. This shows that the above strategy—using the Krichevsky–Trofimov assignment—is optimal up to lower order terms.

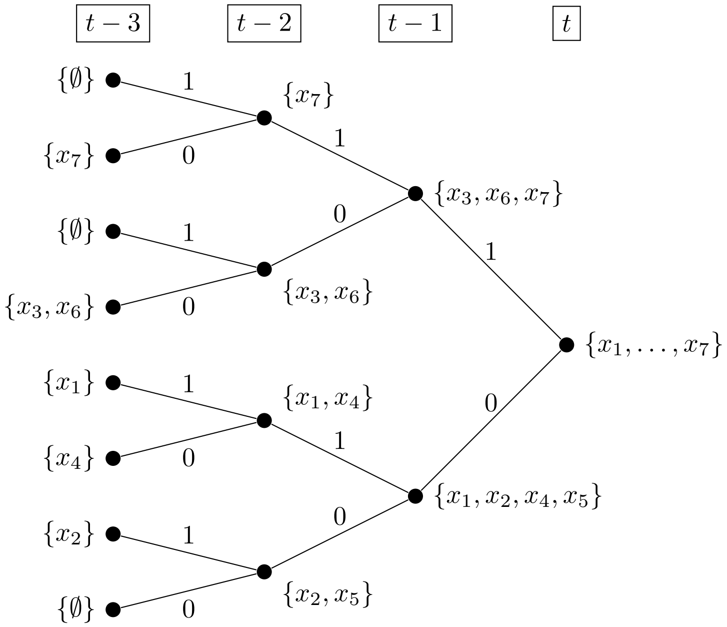
3 Unknown tree

We call a context tree is of depth D if it is a full binary tree of depth D, where node s contains two integers (n_{s0}, n_{s1}), i.e., the number of 0's and 1's in the sequence x^n which occurs right after the context s.

Example 4. Consider the sequence

$$\begin{array}{ccc|cccccccc}
 & & & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & \cdots
\end{array}$$

We can draw the tree, with each node s labeled with $x(s)$:



For example, x_3 is preceded by '001', so is included in $x(\text{'001'})$ (at $t-3$), $x(\text{'01'})$ (at $t-2$) and $x(\text{'1'})$ (at $t-1$). Moreover, for $s = 001$, we have $n_{s0} = n_{s1} = 1$ (corresponding to x_3 and x_6, respectively).

Then the context-tree weighting at any leaf node s is $P_{\text{CTW}}(x^n, s) = P^{\text{KT}}(n_{s0}, n_{s1})$ if $\text{depth}(s) = D$. For any node s other than a leaf node, we have

$$P_{\text{CTW}}(x^n, s) = \frac{1}{2} P^{\text{KT}}(n_{s0}, n_{s1}) + \frac{1}{2} P_{\text{CTW}}(x^n, 0s) P_{\text{CTW}}(x^n, 1s)$$

where $0s$ means the sequence formed by prepending '0' to s, and $1s$ means the sequence formed by prepending '1' to s.

Finally, the CTW probability assignment on x^n can be found recursively at the root via $P_{\text{CTW}}(x^n) = P_{\text{CTW}}(x^n, \emptyset)$.