EE378A Statistical Signal Processing Lecture 10 - 05/04/2017

Lecture 10: Regret Analysis of Context Tree Weighting

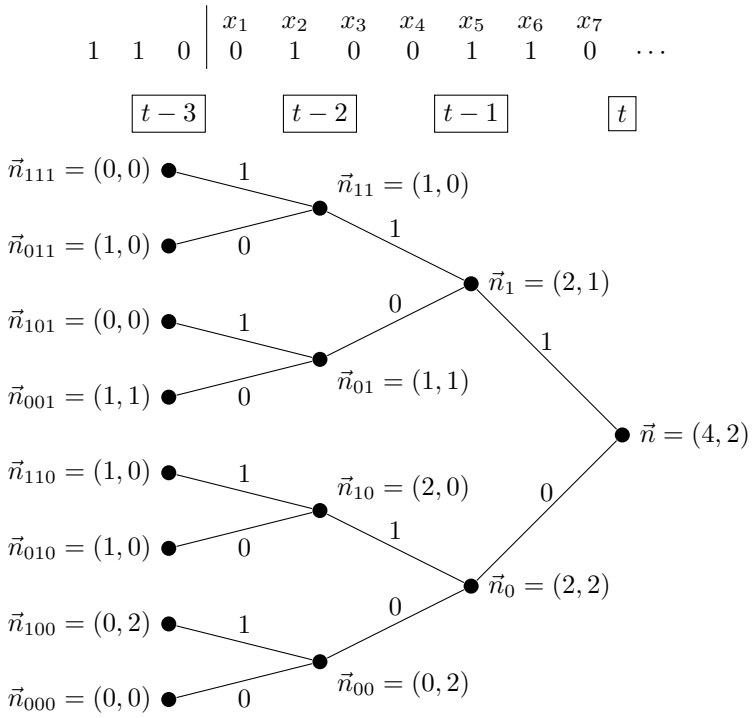
Lecturer: Tsachy Weissman *Scribe: Chuan-Zheng Lee*

In the last lecture we introduced how to use Krichevsky–Trofimov probability assignment to construct the context tree, and the corresponding weighting scheme for tree sources. In this lecture we will analyze the (worst-case, or the average-case) regret of the CTW algorithm under logarithmic loss.

1 Context Tree Weighting: Review

Notation: since the Krichevsky–Trofimov probability assignment only depends on the number of zeros and ones in the sequence, we define $\vec{n} = (n_0(x^n), n_1(x^n))$ and write $P^{\mathrm{KT}}(x^n)$ as $P^{\mathrm{KT}}(\vec{n})$. The CTW algorithm imposes the Krichevsky–Trofimov probability assignment to every context $s \in \mathcal{S}$ separately, and for each context $s \in \mathcal{S}$, we need to associate with a vector $\vec{n}_s = (n_{s0}(x^n), n_{s1}(x^n))$, i.e., the number of occurrences of the context $s0, s1$ in the sequence x^n, respectively.

Now revisit the last example with nodes labeled with their corresponding $\vec{n}$:



The probability assignment of the CTW algorithm on x^n and the context s is given as follows:

$$P_{\mathrm{CTW}}(x^n, s) = \begin{cases} P^{\mathrm{KT}}(\vec{n}_s) & \text{if } \mathbf{depth}(s) = D \\ \frac{1}{2}P^{\mathrm{KT}}(\vec{n}_s) + \frac{1}{2}P_{\mathrm{CTW}}(x^n, 0s)P_{\mathrm{CTW}}(x^n, 1s), & \text{otherwise} \end{cases}.$$

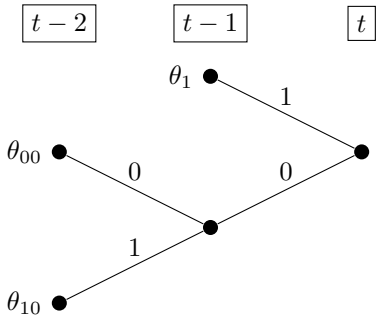
Recursively we find $P_{\mathrm{CTW}}(x^n) = P_{\mathrm{CTW}}(x^n, \emptyset)$, and also the conditional probability

$$P_{\mathrm{CTW}}(x_{n+1}|x^n) = \frac{P_{\mathrm{CTW}}(x^{n+1})}{P_{\mathrm{CTW}}(x^n)}.$$

Remark The choice of the weight $\frac{1}{2}$'s in that second line is not important; it can be changed into any $(\alpha, 1-\alpha)$ with $\alpha \in (0,1)$ with almost no loss in theoretical analysis (as you may see below). In practice, the choice of α depends on the data.

2 Worst-Case Regret of CTW Algorithm

Consider the context tree in the previous example with depth 3, and suppose that the true tree model $\mathcal{S}$ is as follows:



By definition of the CTW algorithm, we have

$$\begin{aligned}
P_{\mathrm{CTW}}(x^n, 00) &\geq \frac{1}{2} P^{\mathrm{KT}}(\vec{n}_{00}) \\
P_{\mathrm{CTW}}(x^n, 10) &\geq \frac{1}{2} P^{\mathrm{KT}}(\vec{n}_{10}) \\
P_{\mathrm{CTW}}(x^n, 1) &\geq \frac{1}{2} P^{\mathrm{KT}}(\vec{n}_{1}) \\
P_{\mathrm{CTW}}(x^n, 0) &\geq \frac{1}{2} P_{\mathrm{CTW}}(x^n, 00) P_{\mathrm{CTW}}(x^n, 10) \\
P_{\mathrm{CTW}}(x^n) &\geq P_{\mathrm{CTW}}(x^n, \emptyset) \\
&\geq \frac{1}{2} P_{\mathrm{CTW}}(x^n, 0) P_{\mathrm{CTW}}(x^n, 1) \\
&\geq \frac{1}{2}\frac{1}{2} P_{\mathrm{CTW}}(x^n, 00) P_{\mathrm{CTW}}(x^n, 10) \frac{1}{2} P^{\mathrm{KT}}(\vec{n}_{1}) \\
&\geq \frac{1}{2^5} P^{\mathrm{KT}}(\vec{n}_{00}) P^{\mathrm{KT}}(\vec{n}_{10}) P^{\mathrm{KT}}(\vec{n}_{1}) \\
&= \frac{1}{2^5} P_{\mathcal{S}}^{\mathrm{KT}}(x^n).
\end{aligned}$$

As a result, for any sequence x^n, in this case we have

$$\begin{aligned}
\log \frac{1}{P_{\mathrm{CTW}}(x^n)} - \min_{\theta} \log \frac{1}{P_{\mathcal{S},\theta}(x^n)} &\leq 5 + \log \frac{1}{P_{\mathcal{S}}^{\mathrm{KT}}(x^n)} - \min_{\theta} \log \frac{1}{P_{\mathcal{S},\theta}(x^n)} \\
&\leq 5 + \frac{3}{2} \log \frac{n}{3} + 3 \\
&= \frac{3}{2} \log \frac{n}{3} + 8.
\end{aligned}$$

In general, repeating the analysis above, we lose a factor of at most $\frac{1}{2}$ for each of the $|\mathcal{S}|$ leaves and $|\mathcal{S}|-1$ internal nodes. Therefore, in general, for any $\mathcal{S}$ corresponding to a tree source of depth $\leq D$,

$$\begin{aligned}
\log \frac{1}{P_{\mathrm{CTW}}(x^n)} &\leq 2|\mathcal{S}| - 1 + \log \frac{1}{P_{\mathcal{S}}^{\mathrm{KT}}(x^n)} \\
&\leq 2|\mathcal{S}| - 1 + \left[\frac{|\mathcal{S}|}{2} \log \frac{n}{|\mathcal{S}|} + |\mathcal{S}| + \min_{\theta} \log \frac{1}{P_{\mathcal{S},\theta}(x^n)} \right] \\
&\leq \frac{|\mathcal{S}|}{2} \log \frac{n}{|\mathcal{S}|} + 3|\mathcal{S}| - 1 + \min_{\theta} \log \frac{1}{P_{\mathcal{S},\theta}(x^n)},
\end{aligned}$$

which again shows the order-optimality of the context-tree weighting for any *unknown* tree model with depth at most D.

3 Average Regret of CTW Algorithm

Recall that the worst-case regret under logarithmic loss is given by

$$\mathrm{WCR}_n(P, \mathcal{F}) = \max_{x^n} \left[\log \frac{1}{P(x^n)} - \min_{Q \in \mathcal{F}} \log \frac{1}{Q(x^n)} \right].$$

Now, for all true probability distribution $Q \in \mathcal{F}$, we have

$$\begin{aligned}
D(Q_{x^n} \| P_{x^n}) &= \mathbb{E}_Q \log \frac{Q(x^n)}{P(x^n)} \\
&\leq \mathbb{E}_Q \left[\log \frac{1}{P(x^n)} - \min_{Q' \in \mathcal{F}} \log \frac{1}{Q'(x^n)} \right] \\
&\leq \mathrm{WCR}_n(P, \mathcal{F})
\end{aligned}$$

which shows that the average-case regret is upper bounded by the worst-case regret:

$$\sup_{Q \in \mathcal{F}} D(Q_{x^n} \| P_{x^n}) \leq \mathrm{WCR}_n(P, \mathcal{F}).$$

Applying this result to the tree source, for any tree source $P_{\mathcal{S},\theta}$ with $\mathrm{depth}(\mathcal{S}) \leq D$, we have

$$D(P_{\mathcal{S},\Theta}^n \| P_{\mathrm{CTW}}^n) \leq \frac{|\mathcal{S}|}{2} \log n + O(1).$$

As a result, we know from Lecture 7 that for prediction under a general loss function Λ,

$$\mathbb{E}\left[L_{F^{P_{\mathrm{CTW}}}}(X^n) - L_{F^{P_{\mathcal{S},\theta}}}(X^n) \right] \leq \Lambda_{\max} \sqrt{\frac{2}{n} \left(\frac{|\mathcal{S}|}{2} \log n + O(1) \right)}.$$