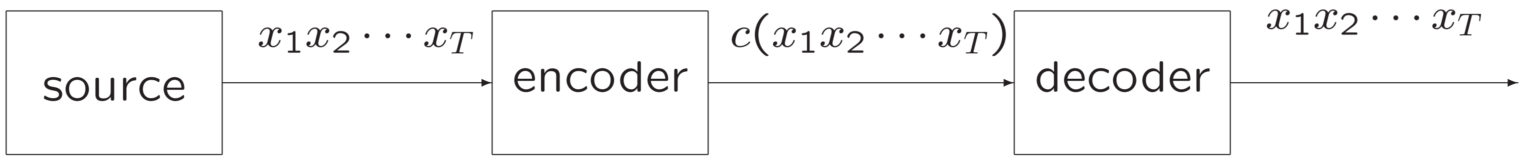
Context-Tree Weighting and Maximizing: Processing Betas

Frans Willems, Tjalling Tjalkens, and Tanya Ignatenko,
Eindhoven University of Technology,
Eindhoven, The Netherlands

I. Noiseless Source Coding

The *source* produces a *sequence* $x_1^T = x_1 x_2 \cdots x_T$ with symbols $x_t \in \{0, 1\}$ for $t = 1, T$ with actual probability $P_a(x_1^T)$.



Example: Independent identically distributed (I.I.D.) source with parameter θ. Let

$$\begin{aligned} P_a(1) &= \theta, \text{ and} \\ P_a(0) &= 1 - \theta, \end{aligned}$$

for some $0 \leq \theta \leq 1$. Then a sequence x^T containing a zeros and b ones has

$$P_a(x_1^T) = (1 - \theta)^a \theta^b.$$

II. Arithmetic Codes

An *arithmetic source code* assigns to source sequence x_1^T a binary codeword $c(x_1^T)$ of length $L(x_1^T)$. Arithmetic coding achieves

$$L(x_1^T) < \log_2 \frac{1}{P_a(x_1^T)} + 2.$$

Arithmetic codes are *prefix codes*. In a prefix code no codeword is the prefix of any other codeword (prefix condition). $\Rightarrow$ instantaneous decodability.

Example: I.I.D. source with $\theta = 0.2$ and $T = 2$.

x_1^T	$c(x_1^T)$
00	0
01	1011
10	1101
11	11111

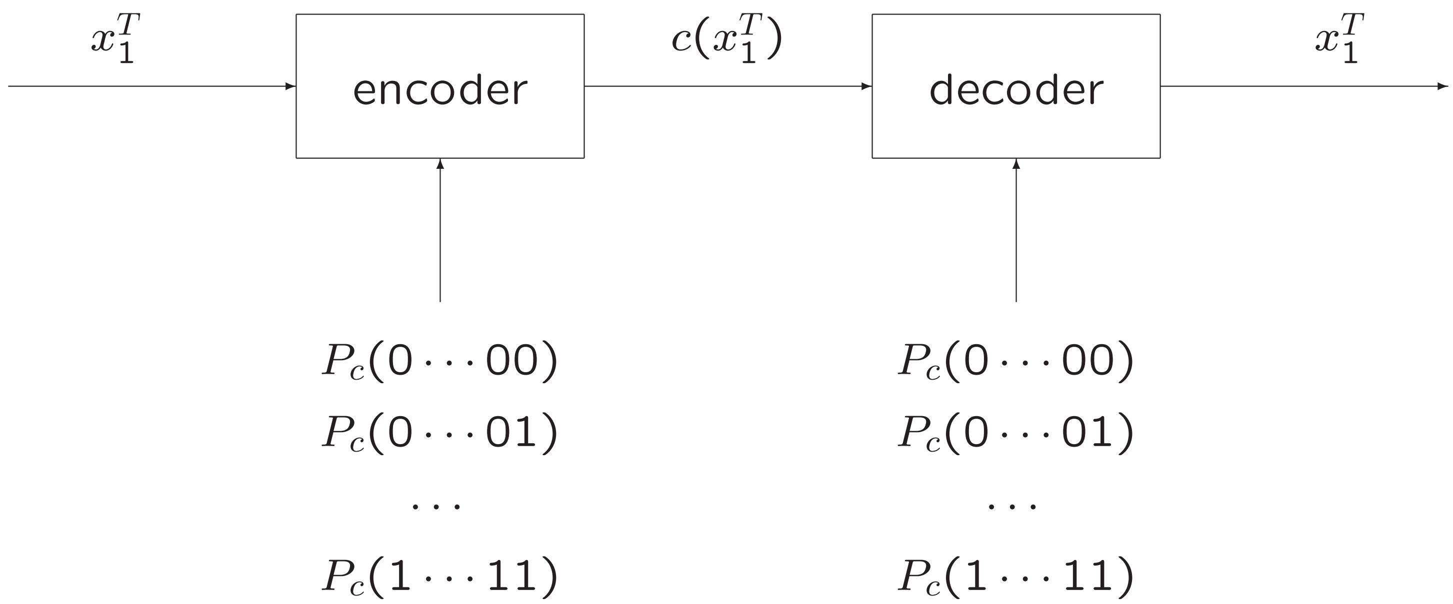
III. Universal Coding

What should we do if the actual probabilities $P_a(x_1^T)$ are *not known*?

Arithmetic coding is still possible if instead of $P_a(x_1^T)$ we use *coding probabilities* $P_c(x_1^T)$ satisfying

$$\begin{aligned} P_c(x_1^T) &> 0 \text{ for all } x_1^T\text{, and} \\ \sum_{x_1^T} P_c(x_1^T) &= 1. \end{aligned}$$

System:



IV. Individual Redundancy

The *individual redundancy* $\rho(x_1^T)$ of a sequence x_1^T is defined as the codeword-length minus *ideal* codeword-length, i.e.

$$\begin{aligned}
\rho(x_1^T) &= L(x_1^T) - \log_2 \frac{1}{P_a(x_1^T)} \\
&< \log_2 \frac{1}{P_c(x_1^T)} + 2 - \log_2 \frac{1}{P_a(x_1^T)} \\
&= \log_2 \frac{P_a(x_1^T)}{P_c(x_1^T)} + 2,
\end{aligned}$$

PROBLEM : How do we choose the coding probabilities $P_c(x_1^T)$?

V. I.I.D. Source with Unknown θ

A good coding probability for a sequence x_1^T that contains a zeroes and b ones is

$$P_e(a,b) \triangleq \frac{\frac{1}{2} \cdot \frac{3}{2} \cdot \ldots \cdot (a - \frac{1}{2}) \cdot \frac{1}{2} \cdot \frac{3}{2} \cdot \ldots \cdot (b - \frac{1}{2})}{1 \cdot 2 \cdot \ldots \cdot (a+b)}$$

(Krichevsky-Trofimov estimator).

Properties:

- Upperbound:

$$\log_2 \frac{P_a(x_1^T)}{P_c(x_1^T)} = \log_2 \frac{\theta^a (1-\theta)^b}{P_e(a,b)} \leq \frac{1}{2} \log_2 T + 1.$$

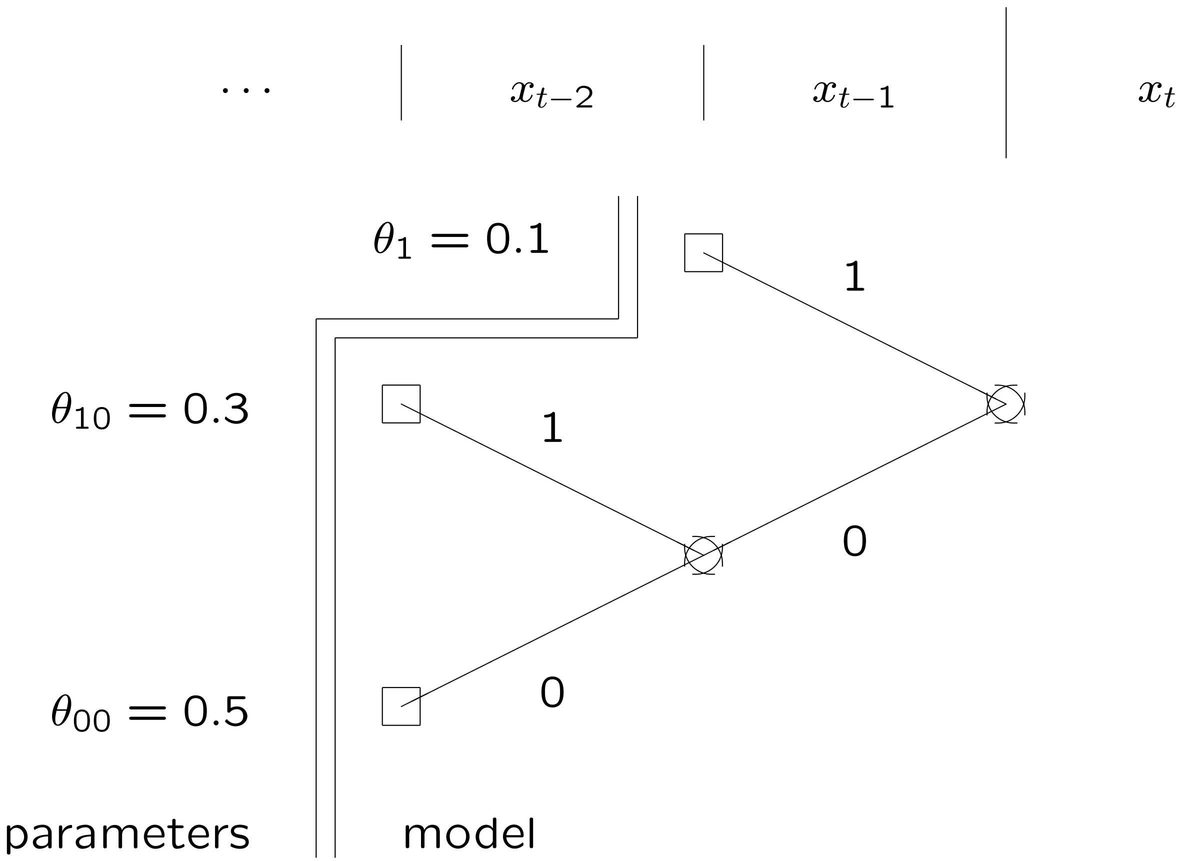
  for all θ and x_1^T with a zeros and b ones.

- Asymptotically optimal (achieves Rissanen's lower bound).

- Recursive behavior:

$$P_e(a+1,b) = \frac{a+1/2}{a+b+1} \cdot P_e(a,b).$$

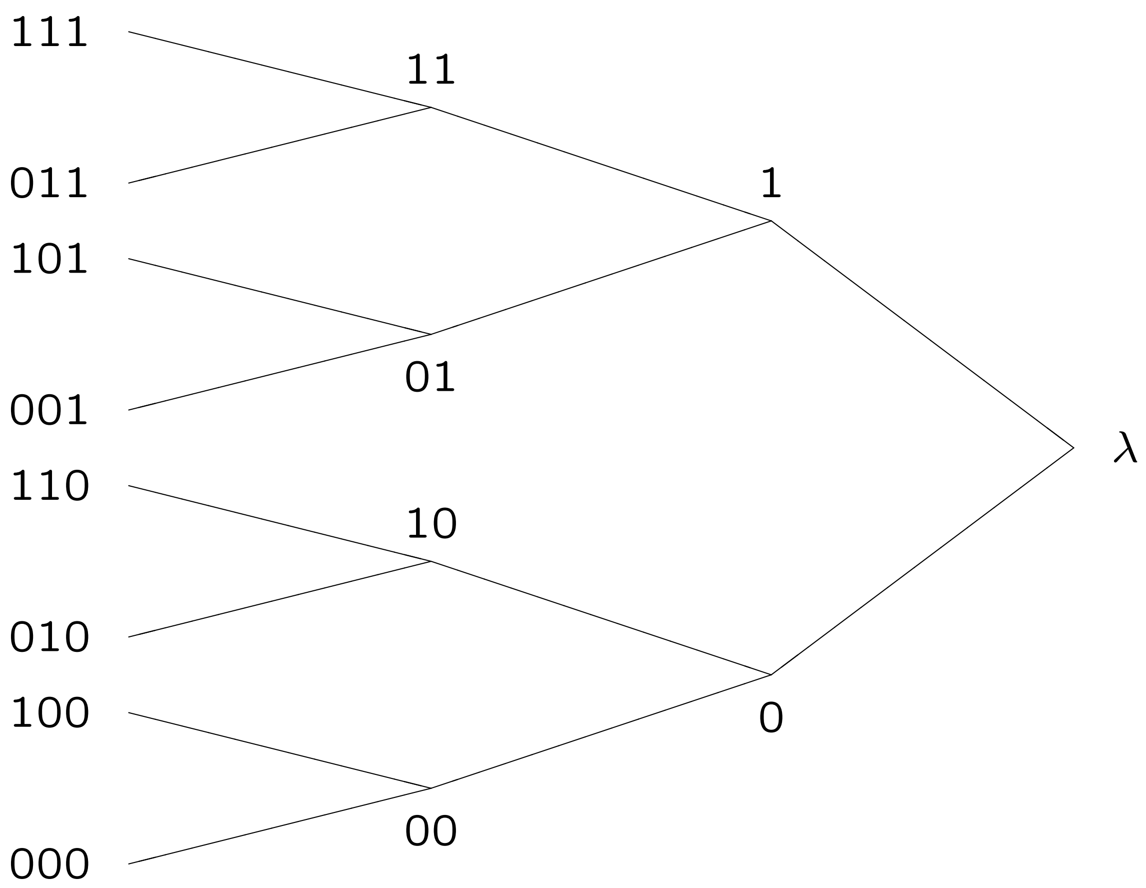
VI. Binary Tree Sources (Example)



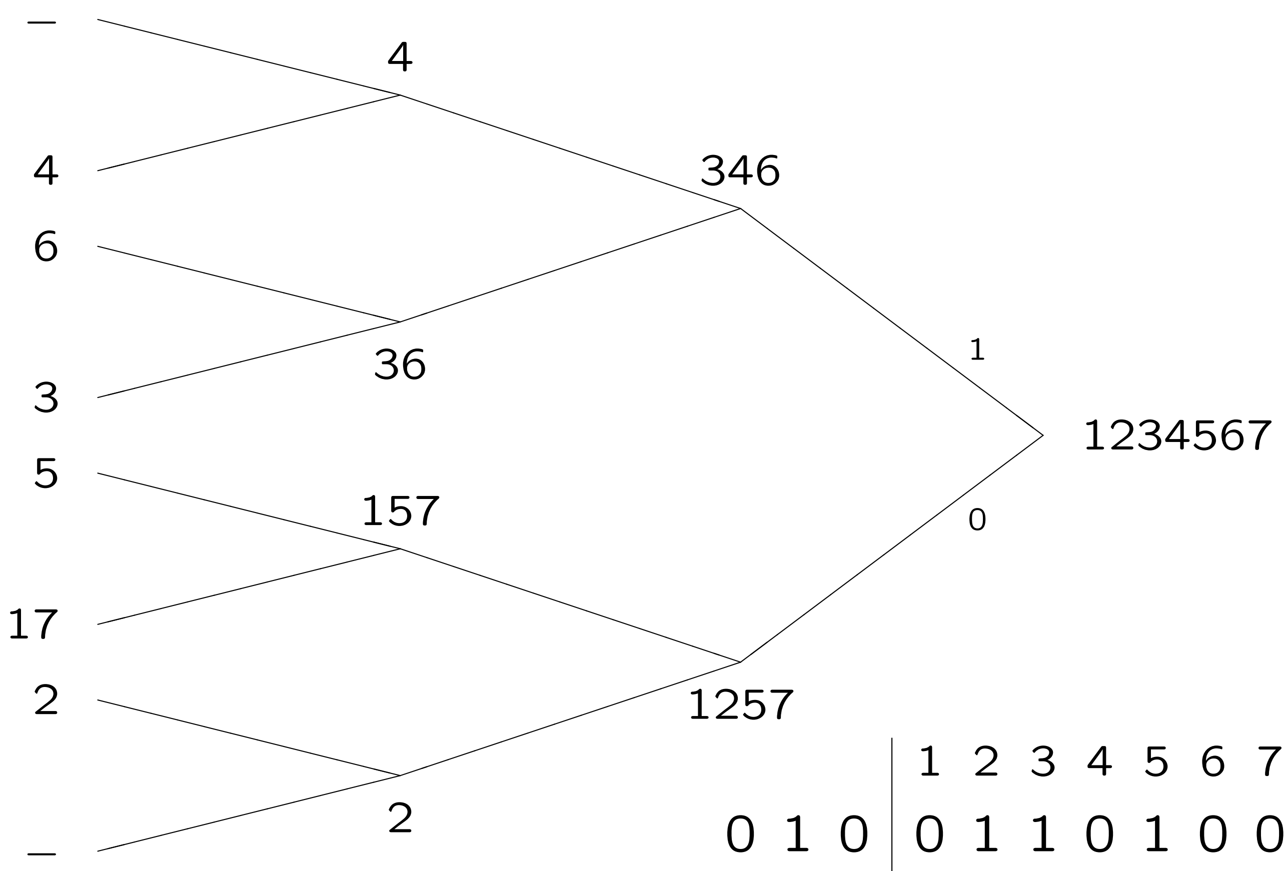
$$
\begin{aligned}
P_a(X_t = 1|\cdots, X_{t-1} = 1) &= 0.1 \\
P_a(X_t = 1|\cdots, X_{t-2} = 1, X_{t-1} = 0) &= 0.3 \\
P_a(X_t = 1|\cdots, X_{t-2} = 0, X_{t-1} = 0) &= 0.5
\end{aligned}
$$

VII. Context-Tree Weighting

A *context tree* is a tree-like data-structure with depth D. Node s contains the sequence of source symbols that have occurred following context s.



Context-tree *splits up* sequences in subsequences.



	1	2	3	4	5	6	7
0 1 0	0	1	1	0	1	0	0
past	x_1^T						

Recursive weighting (WST 1995) yields the coding probability:

$$
\begin{aligned}
P_w^s &\triangleq P_e(a_s, b_s) \text{ for } s \text{ at level } D, \\
P_w^s &\triangleq \frac{P_e(a_s, b_s) + P_w^{0s} \cdot P_w^{1s}}{2} \text{ for } s \text{ elsewhere.}
\end{aligned}
$$

for the subsequence that corresponds to node s.
In the *root* λ of the context-tree the coding probability P_w^λ for the entire source sequence x_1^T.

Total individual redundancy:

$$
\rho(x_1^T) < \Gamma_D(\mathcal{S}) + \left(\frac{|\mathcal{S}|}{2} \log_2 \frac{T}{|\mathcal{S}|} + |\mathcal{S}| \right) + 2 \text{ bits},
$$

where

$$
\Gamma_D(\mathcal{S}) \triangleq 2|\mathcal{S}| - 1 - |\{s \in \mathcal{S}, \text{depth}(s) = D\}|.
$$

Asymptotically optimal (achieves Rissanen's lower bound).

CTW is a Bayes method:

It implements a "weighting" over all tree-models with depth not exceeding D, i.e.

$$P_w^\lambda = \sum_{\mathcal{S} \in \mathcal{T}_\mathcal{D}} P(\mathcal{S}) P_e(x_1^T | \mathcal{S}),$$

with

$$P_e(x_1^T | \mathcal{S}) = \Pi_{s \in \mathcal{S}} P_e(a_s, b_s),$$

and *a priori* tree-model probability

$$P(\mathcal{S}) = 2^{-\Gamma_D(\mathcal{S})}.$$

VIII. Context-Tree Maximizing

The CTW-method is a *one-pass algorithm*, code digits are produced "on the fly". In a *two-pass system* the entire source sequence x_1^T is observed first, and after that a codeword is constructed. Consider the following *two-pass method*:

- After observing x_1^T determine the "best" tree model $\widehat{\mathcal{S}}$.
- Encode this model $\widehat{\mathcal{S}}$, for this $\log_2 \frac{1}{P(\widehat{\mathcal{S}})} = \Gamma_D(\widehat{\mathcal{S}})$ binary digits are needed.
- Encode the sequence x_1^T given this model $\widehat{\mathcal{S}}$ using $< \log_2 \frac{1}{P_e(x_1^T|\widehat{\mathcal{S}})} + 2$ binary digits.

The CTM method chooses, given x_1^T, the model $\widehat{\mathcal{S}}$ that maximizes over $\mathcal{T_D}$ the product

$$P(\widehat{\mathcal{S}})P_e(x_1^T|\widehat{\mathcal{S}}) = 2^{-\Gamma_D(\widehat{\mathcal{S}})} \cdot P_e(x_1^T|\widehat{\mathcal{S}}),$$

and thereby minimizes the total codeword length.

This is done recursively, using a context-tree, by setting

$$P_m^s = P_e(a_s, b_s) \text{ for } s \text{ at level } D,$$
$$P_m^s = \frac{\max[P_e(a_s, b_s), P_m^{0s} \cdot P_m^{1s}]}{2} \text{ for } s \text{ elsewhere.}$$

We assume that the entire sequence x_1^T was already processed in the context tree.
We will find the best model $\widehat{\mathcal{S}}$ by tracking the maximization procedure, starting in the *root* λ of the context-tree. (WST 1993, Nohre 1994)

Total individual redundancy:
CTM achieves exactly the same upper bounds on the individual redundancy as the CTW method. In practice CTW achieves better results though.

IX. Betas: Introduction

Consider an internal node s in the context tree $\mathcal{T}_{\mathcal{D}}$ and the corresponding *conditional* weighted probability $P_w^s(X_t = 1|x_1^{t-1})$. Assuming that $0s$ (and not $1s$) is a suffix of the context x_{1-D}^0, x_1^{t-1} of x_t, we obtain for this probability that

$$P_w^s(X_t = 1|x_1^{t-1}) = \frac{P_e^s(x_1^{t-1}, X_t = 1) + P_w^{0s}(x_1^{t-1}, X_t = 1)P_w^{1s}(x_1^{t-1})}{P_e^s(x_1^{t-1}) + P_w^{0s}(x_1^{t-1})P_w^{1s}(x_1^{t-1})}$$

$$= \frac{\beta^s(x_1^{t-1})P_e^s(X_t = 1|x_1^{t-1}) + P_w^{0s}(X_t = 1|x_1^{t-1})}{\beta^s(x_1^{t-1}) + 1} \quad (1)$$

where

$$\beta^s(x_1^{t-1}) \triangleq \frac{P_e^s(x_1^{t-1})}{P_w^{0s}(x_1^{t-1})P_w^{1s}(x_1^{t-1})}. \quad (2)$$

If we start in the context-leaf and work our way down to the root, we finally find $P_w^\lambda(X_t = 1|x_1^{t-1})$.

Implementation

Assume that in node s the counts $a_s(x_1^{t-1})$ and $b_s(x_1^{t-1})$ are stored, as well as $\beta^s(x_1^{t-1})$. We then get the following sequence of operations:

1. Node $0s$ delivers cond. wei. probability $P_w^{0s}(X_t = 1|x_1^{t-1})$ to node s.
2. Cond. est. probability $P_e^s(X_t = 1|x_1^{t-1})$ is determined as follows:

$$P_e^s(X_t = 1|x_1^{t-1}) = \frac{b_s(x_1^{t-1}) + 1/2}{a_s(x_1^{t-1}) + b_s(x_1^{t-1}) + 1}. \tag{3}$$

3. Now $P_w^s(X_t = 1|x_1^{t-1})$ can be computed as in (1).
4. The ratio $\beta^s(\cdot)$ is then updated with symbol x_t as follows:

$$\beta^s(x_1^{t-1}, x_t) = \beta^s(x_1^{t-1}) \cdot \frac{P_e^s(X_t = x_t|x_1^{t-1})}{P_w^{0s}(X_t = x_t|x_1^{t-1})}. \tag{4}$$

5. Finally, depending on the value x_t, either count $a_s(x_1^{t-1})$ or $b_s(x_1^{t-1})$ is incremented.

X. Betas: A Posteriori Tree-Model Probs.

Consider tree-model $\mathcal{S}$ and let $\mathcal{S}_s$ be its sub-tree rooted at s. Define the conditional probability of the sub-tree $\mathcal{S}_s$ given x_1^T as

$$Q_w^s(\mathcal{S}_s) \triangleq \frac{2^{-\Gamma_D(\mathcal{S}_s)} \prod_{s \in \mathcal{S}_s} P_e(a_s, b_s)}{P_w^s},$$

where the cost of sub-model $\mathcal{S}_s$ is defined as

$$\Gamma_D(\mathcal{S}_s) \triangleq 2|\mathcal{S}_s| - 1 - |\{s \in \mathcal{S}_s, \text{depth}(s) = D\}|.$$

If $|\mathcal{S}_s| > 1$ node s can not be at level D and we can split up $\mathcal{S}_s$ into a sub-model $\mathcal{S}_{0s}$ and a sub-model $\mathcal{S}_{1s}$ and we obtain:

$$\begin{aligned} Q_w^s(\mathcal{S}_s) &= \frac{2^{-\Gamma_D(\mathcal{S}_{0s})} \prod_{s \in \mathcal{S}_{0s}} P_e(a_s, b_s)}{P_w^{0s}} \cdot \frac{2^{-\Gamma_D(\mathcal{S}_{1s})} \prod_{s \in \mathcal{S}_{1s}} P_e(a_s, b_s)}{P_w^{1s}} \\ &\quad \cdot \frac{P_w^{0s} P_w^{1s}}{P_e(a_s, b_s) + P_w^{0s} P_w^{1s}} \\ &= Q_w^{0s}(\mathcal{S}_{0s}) Q_w^{1s}(\mathcal{S}_{1s}) \frac{1}{\beta_s + 1}. \end{aligned}$$

When sub-model $\mathcal{S}_s$ contains only one leaf-node s, not at depth D, then

$$Q_w^s(\mathcal{S}_s) = \frac{P_e(a_s, b_s)}{P_e(a_s, b_s) + P_w^{0s} P_w^{1s}} = \frac{\beta_s}{\beta_s + 1}.$$

If sub-model $\mathcal{S}_s$ consists only of a single leaf-node s at level D then

$$Q_w^s(\mathcal{S}_s) = 1.$$

Summarizing the three considered cases we can write

$$Q_w^s(\mathcal{S}_s) = \begin{cases} Q_w^{0s}(\mathcal{S}_{0s}) Q_w^{1s}(\mathcal{S}_{1s}) \frac{1}{\beta_s + 1} & \text{if } |\mathcal{S}_s| > 1, \\ \frac{\beta_s}{\beta_s+1} \text{ if depth}(s) < D & \text{for } |\mathcal{S}_s| = 1, \\ 1 \text{ if depth}(s) = D & \text{for } |\mathcal{S}_s| = 1. \end{cases}$$

For the tree model $\mathcal{S}$ (rooted in λ), we can write for its *a posteriori probability* after having observed x_1^T that

$$P_w(\mathcal{S}|x_1^T) = \frac{2^{-\Gamma_D(\mathcal{S})} \prod_{s \in \mathcal{S}} P_e(a_s, b_s)}{P_w^\lambda} = Q_w^\lambda(\mathcal{S}).$$

XI. Betas: Finding the MAP Tree-Model

The CTM method produces the MAP tree-model given the source sequence x_1^T. We want to determine the MAP-model based on the β's in the *weighted* context-tree.

First we compute the probability of the MAP sub-model corresponding to a node s at depth $< D$. For such a node

$$\max_{\mathcal{S}_s} Q_w^s(\mathcal{S}_s) = \max[\frac{1}{\beta_s+1} \max_{\mathcal{S}_{0s}} Q_w^{0s}(\mathcal{S}_{0s}) \max_{\mathcal{S}_{1s}} Q_w^{1s}(\mathcal{S}_{1s}), \frac{\beta_s}{\beta_s+1}].$$

The last term corresponds to the sub-model which has only a single leaf-node at s. The first term to all larger sub-models.

For a node at depth D only the one-leaf sub-model plays a role and

$$\max_{\mathcal{S}_s} Q_w^s(\mathcal{S}_s) = 1.$$

If we now define for all nodes $s \in \mathcal{T}_\mathcal{D}$ the MAP sub-model probability

$$Q^s_{mw} \triangleq \max_{\mathcal{S}_s} Q^s_w(\mathcal{S}_s),$$

then the following recursive equation holds:

$$Q^s_{mw} = \begin{cases} \max[Q^{0s}_{mw} Q^{1s}_{mw} \frac{1}{\beta_s+1}, \frac{\beta_s}{\beta_s+1}] & \text{if depth}(s) < D, \\ 1 & \text{if depth}(s) = D. \end{cases}$$

Now in the root λ of the context tree we find the maximum a posteriori model probability Q^λ_{mw}. Tracking the procedure, starting in the root of the context tree, yields the MAP-model.

XII. Betas: A Convex Combination

Equation (1) expresses the conditional weighted probability of the root as a linear combination of the estimated probabilities of the nodes along the context path $x_{t-D}x_{t-D+1}\cdots x_{t-1}$. This results in

$$P_w^\lambda(X_t = 1|x_1^{t-1}) = \sum_{d=0,D} \mu^{s_d}(x_1^{t-1}) P_e^s(X_t = 1|x_1^{t-1}) \tag{5}$$

where $s_0 \triangleq \lambda$ and $s_d \triangleq x_{t-d}\cdots x_{t-1}$ for $d = 1, \cdots, D$, and

$$\mu^{s_d}(x_1^{t-1}) = \frac{\beta^{s_d}(x_1^{t-1})}{\beta^{s_d}(x_1^{t-1})+1} \prod_{i=0,d-1} \frac{1}{\beta^{s_i}(x_1^{t-1})+1}, \tag{6}$$

for $d = 0, 1, \cdots, D-1$ and

$$\mu^{s_D}(x_1^{t-1}) = \prod_{i=0,D-1} \frac{1}{\beta^{s_i}(x_1^{t-1})+1}. \tag{7}$$

If we observe that $\mu^{s_d}(x_1^{t-1}) \geq 0$ for $d = 0, 1, \cdots, D$ and

$$\sum_{d=0,D} \mu^{s_d}(x_1^{t-1}) = 1,$$

we may conclude that (5) is actually a convex combination.

XIII. Betas: A Posteriori Node Probabilities

We can define the a posteriori *node* probability of a node $s \in \mathcal{T}_{\mathcal{D}}$ as

$$Q_w(s) \triangleq \sum_{\mathcal{S}: s \in \mathcal{S}} Q_w^{\lambda}(\mathcal{S}),$$

where the summation is over all models $\mathcal{S}$ that contain leaf s.

It now can be shown that for all $s \in \mathcal{T}_{\mathcal{D}}$

$$\mu^s = Q_w(s), \tag{8}$$

where μ_s is as in (6) and (7).

XIV. Betas: Difference CTW and CTM

Let $\hat{\mathcal{S}}$ be the MAP model, then

$$1 \geq \frac{2^{-\Gamma_D(\hat{\mathcal{S}})} \prod_{s \in \hat{\mathcal{S}}} P_e(a_s, b_s)}{P_w^\lambda} = \frac{P_m^\lambda}{P_w^\lambda} = Q_w^\lambda(\hat{\mathcal{S}}) = Q_{mw}^\lambda.$$

For the difference in codeword lengths for CTW and CTM we can write

$$L_w(x_1^T) - L_m(x_1^T) = \lceil \log_2 \frac{1}{P_w^\lambda(x_1^T)} \rceil + 1 - \lceil \log_2 \frac{1}{P_m^\lambda(x_1^T)} \rceil - 1 \leq 0.$$

However we can also show that

$$\begin{aligned} L_w(x_1^T) - L_m(x_1^T) &< \log_2 \frac{1}{P_w^\lambda(x_1^T)} + 2 - \log_2 \frac{1}{P_m^\lambda(x_1^T)} - 1 \\ &= \log_2 Q_{mw}^\lambda + 1, \end{aligned}$$

and similarly

$$L_w(x_1^T) - L_m(x_1^T) > \log_2 Q_{mw}^\lambda - 1.$$

XV. Conclusion

- Betas simplify the implementation.

- Based on betas we can compute:

  - A posteriori probabilities,

  - MAP tree-model,

  - $P_w^\lambda(X_t = 1|x_1^{t-1})$ as convex combination of cond. estim. probabilities along context path,

  - difference between CTW and CTM codeword lengths.

- Similar results hold for weightings other than $(1/2, 1/2)$.