Improvements to the Sequence Memoizer Supplementary Material

Jan Gasthaus
Gatsby Computational Neuroscience Unit
University College London
London, WC1N 3AR, UK
j.gasthaus@gatsby.ucl.ac.uk

Yee Whye Teh
Gatsby Computational Neuroscience Unit
University College London
London, WC1N 3AR, UK
ywteh@gatsby.ucl.ac.uk

1 Generalized Stirling Numbers of Type $(-1, -d, 0)$

In Appendix A.5 of [1] it was shown by induction that

$$\sum_{A \in \mathcal{A}_{ct}} \prod_{a \in A} [1-d]_1^{|a|-1} = S_d(c,t) \tag{1}$$

where $S_d(c,t)$ is a generalized Stirling number of type $(-1,-d,0)$ [2]. These can be computed recursively as follows:

$$S_d(1,1) = S_d(0,0) = 1 \tag{2}$$
$$S_d(c,0) = S_d(0,t) = 0 \qquad \text{for } c,t > 0 \tag{3}$$
$$S_d(c,t) = 0 \qquad \text{for } t > c \tag{4}$$
$$S_d(c,t) = S_d(c-1,t-1) + (c-1-dt) S_d(c-1,t) \qquad \text{for } 0 < t \leq c \tag{5}$$

2 More Verbose Proof of Theorem 1

Theorem 1. *Suppose $A_2 \in \mathcal{A}_c$, $A_1 \in \mathcal{A}_{|A_2|}$, $C \in \mathcal{A}_c$ and $F_a \in \mathcal{A}_{|a|}$ for each $a \in C$ are related as above. Then the following describe equivalent distributions:*
(I) $A_2 \sim \mathrm{CRP}_c(\alpha d_2, d_2)$ and $A_1|A_2 \sim \mathrm{CRP}_{|A_2|}(\alpha, d_1)$.
(II) $C \sim \mathrm{CRP}_c(\alpha d_2, d_1 d_2)$ and $F_a|C \sim \mathrm{CRP}_{|a|}(-d_1 d_2, d_2)$ for each $a \in C$.

Proof. In the proof we will use the identity $[\beta\delta + \delta]_\delta^{n-1} = \delta^{n-1}[\beta+1]_1^{n-1}$ (for all β, δ, n) several times; let us call it Identity 1.

To complete the proof, we simply show that the joint distributions are the same. Starting with the definition of the the CRP distribution

$$P(A) = \frac{[\alpha+d]_d^{|A|-1}}{[\alpha+1]_1^{c-1}} \prod_{a \in A} [1-d]_1^{|a|-1} \quad \text{for each } A \in \mathcal{A}_c, \tag{6}$$

we have the following by multiplying the two distributions together:

$$P(A_1, A_2) = \left(\frac{[\alpha+d_1]_{d_1}^{|A_1|-1}}{[\alpha+1]_1^{|A_2|-1}} \prod_{a \in A_1} [1-d_1]_1^{|a|-1} \right) \left(\frac{[\alpha d_2 + d_2]_{d_2}^{|A_2|-1}}{[\alpha d_2 + 1]_1^{c-1}} \prod_{b \in A_2} [1-d_2]_1^{|b|-1} \right) \tag{7}$$

Re-arranging terms yields:

$$= \frac{[\alpha+d_1]_{d_1}^{|A_1|-1}}{[\alpha+1]_1^{|A_2|-1}} \frac{[\alpha d_2 + d_2]_{d_2}^{|A_2|-1}}{[\alpha d_2 + 1]_1^{c-1}} \left(\prod_{a \in A_1} [1-d_1]_1^{|a|-1} \right) \left(\prod_{b \in A_2} [1-d_2]_1^{|b|-1} \right) \tag{8}$$

Using Identity 1 on $[\alpha d_2 + d_2]_{d_2}^{|A_2|-1}$ yields

$$= \frac{[\alpha + d_1]_{d_1}^{|A_1|-1}}{[\alpha+1]_1^{|A_2|-1}} \frac{d_2^{|A_2|-1}[\alpha+1]_1^{|A_2|-1}}{[\alpha d_2 + 1]_1^{c-1}} \Big(\prod_{a \in A_1} [1-d_1]_1^{|a|-1} \Big) \Big(\prod_{b \in A_2} [1-d_2]_1^{|b|-1} \Big) \tag{9}$$

Cancelling one of the resulting terms we get:

$$= \frac{[\alpha + d_1]_{d_1}^{|A_1|-1}}{} \frac{d_2^{|A_2|-1}}{[\alpha d_2 + 1]_1^{c-1}} \Big(\prod_{a \in A_1} [1-d_1]_1^{|a|-1} \Big) \Big(\prod_{b \in A_2} [1-d_2]_1^{|b|-1} \Big) \tag{10}$$

Multiplying and dividing by $d_2^{|a|-1}$ within the first product and using Identity 1 again we get:

$$= \frac{[\alpha + d_1]_{d_1}^{|A_1|-1}}{} \frac{d_2^{|A_2|-1}}{[\alpha d_2 + 1]_1^{c-1}} \Big(\prod_{a \in A_1} \frac{1}{d_2^{|a|-1}} [d_2 - d_1 d_2]_{d_2}^{|a|-1} \Big) \Big(\prod_{b \in A_2} [1-d_2]_1^{|b|-1} \Big) \tag{11}$$

Using $\sum_{a \in A_1} |a| = |A_2|$ to take the $1/(d_2^{|a|-1})$ term out of the first product:

$$= \frac{[\alpha + d_1]_{d_1}^{|A_1|-1}}{} \frac{d_2^{|A_2|-1}}{[\alpha d_2 + 1]_1^{c-1}} \frac{1}{d_2^{|A_2|-|A_1|}} \Big(\prod_{a \in A_1} [d_2 - d_1 d_2]_{d_2}^{|a|-1} \Big) \Big(\prod_{b \in A_2} [1-d_2]_1^{|b|-1} \Big) \tag{12}$$

Using $d_2^{|A_2|-|A_1|} = d_2^{|A_2|-1} d_2^{-|A_1|+1}$ and cancelling the $d_2^{|A_2|-1}$ term:

$$= \frac{[\alpha + d_1]_{d_1}^{|A_1|-1}}{} \frac{}{[\alpha d_2 + 1]_1^{c-1}} \frac{1}{d_2^{-|A_1|+1}} \Big(\prod_{a \in A_1} [d_2 - d_1 d_2]_{d_2}^{|a|-1} \Big) \Big(\prod_{b \in A_2} [1-d_2]_1^{|b|-1} \Big) \tag{13}$$

Multiplying and dividing by $d_2^{|A_1|-1}$ and using Identity 1 a third time we get:

$$= \frac{[\alpha d_2 + d_1 d_2]_{d_1 d_2}^{|A_1|-1}}{d_2^{|A_1|-1}} \frac{}{[\alpha d_2 + 1]_1^{c-1}} \frac{1}{d_2^{-|A_1|+1}} \Big(\prod_{a \in A_1} [d_2 - d_1 d_2]_{d_2}^{|a|-1} \Big) \Big(\prod_{b \in A_2} [1-d_2]_1^{|b|-1} \Big) \tag{14}$$

Finally, cancelling terms again we get:

$$= \frac{[\alpha d_2 + d_1 d_2]_{d_1 d_2}^{|A_1|-1}}{[\alpha d_2 + 1]_1^{c-1}} \Big(\prod_{a \in A_1} [d_2 - d_1 d_2]_{d_2}^{|a|-1} \Big) \Big(\prod_{b \in A_2} [1-d_2]_1^{|b|-1} \Big) \tag{15}$$

Re-grouping the products and expressing the same quantities in terms of C and $\{F_a\}$,

$$= \frac{[\alpha d_2 + d_1 d_2]_{d_1 d_2}^{|C|-1}}{[\alpha d_2 + 1]_1^{c-1}} \prod_{a \in C} \Big([d_2 - d_1 d_2]_{d_2}^{|F_a|-1} \prod_{b \in F_a} [1-d_2]_1^{|b|-1} \Big) = P(C, \{F_a\}_{a \in C}) \tag{16}$$

By comparison with (6), we see that conditioning on C each $F_a \sim \mathrm{CRP}_{|a|}(-d_1 d_2, d_2)$.

Marginalizing $\{F_a\}$ out using the normalization constant from (6) we have:

$$P(C) = \frac{[\alpha d_2 + d_1 d_2]_{d_1 d_2}^{|C|-1}}{[\alpha d_2 + 1]_1^{c-1}} \prod_{a \in C} [1 - d_1 d_2]_1^{|a|-1} \tag{17}$$

So $C \sim \mathrm{CRP}_c(\alpha d_2, d_1 d_2)$ and *(I)*⇒*(II)*. Reversing the same argument shows that *(II)*⇒*(I)*. □

References

[1] Y. W. Teh. A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore, 2006.

[2] L. C. Hsu and P. J.-S. Shiue. A unified approach to generalized Stirling numbers. *Advances in Applied Mathematics*, 20:366–384, 1998.