Supplement: A Discriminative Latent Variable Model for Online Clustering

Rajhans Samdani, Google Research
Kai-Wei Chang, University of Illinois
Dan Roth, University of Illinois

This supplement provides the proof that the probability of clustering as per the left-linking tree model is the same as the probability of a clustering as per the L³M model and also provides the results with randomized ordering of items discussed in Section 5.4.

1. Proof of Theorem 1

The main paper (Eq. 4) presents the probability of a clustering C as per the left-linking tree model as:

\[ Pr[C; d, w] = \sum_{\sigma \in \mathcal{C}} Pr[j \leftarrow i; d, w|C(i, j) = Z(C; d, w, \gamma) / Z(d, w, \gamma), \]

where \( Z(d, w, \gamma) = \prod_{i=1}^m Z_i(d, w, \gamma) \) is the partition function and \( Z(C; d, w, \gamma) = \prod_{i=1}^m Z_i(C; d, w, \gamma) \).

For the left-linking tree model, the probability of a left-linking tree z is represented as

\[ Pr[z; d, w] = \frac{1}{T(d, w, \gamma)} \exp \left( \frac{1}{\gamma} \sum_{i,j \in z} w \cdot \phi(i, j) \right), \]

where \( T(d, w, \gamma) = \sum_{z \in \mathcal{Z}_d} \exp \left( \frac{1}{\gamma} \sum_{i,j \in z} w \cdot \phi(i, j) \right) \) is the left-linking tree partition function. The probability of a clustering C as per the left-linking tree model is expressed (Eq. 5 in the main paper) as the sum of the probabilities of all the left-linking trees consistent with C:

\[ Pr[C; d, w] = \sum_{z \in \mathcal{Z}_d} Pr[z; d, w] \]

\[ = \frac{1}{T(d, w, \gamma)} \sum_{z \in \mathcal{Z}_d} \exp \left( \frac{1}{\gamma} \sum_{i,j \in z} w \cdot \phi(i, j) \right). \]

Now, following is the theorem stated in the paper.

**Theorem 1.** The probability of a clustering as per the left-linking tree model, expressed in Eq. (2), is the same as the probability of clustering for L³M as expressed in Eq. (1), i.e. \( Pr[C; d, w] = Pr[C; d, w] \).

**Proof.** We will focus on the proof with \( \gamma > 0 \). As the functions in Eq. (1) and Eq. (2) are bounded and continuous, we shall see that the same result will hold for \( \gamma \to 0 \).

First we will prove that the two partition functions are the same: \( T(d, w, \gamma) = Z(d, w, \gamma) \). The proof for the equivalence of the numerators for Eq. (1) and Eq. (2) will be analogous.

Below, we prove \( T(d, w, \gamma) = Z(d, w, \gamma) \) by induction on the number of items, \( m_d \). We abuse the notation and write \( Z(n, w, \gamma) \) as the partition function for L³M when considering only the first \( n \) items. We use the notation \( T(n, w, \gamma) \) similarly.

**Base case:** \( m_d = 1 \). With just one actual item and one dummy item 0, \( Z(1, w, \gamma) = \exp(\frac{1}{\gamma}(w \cdot \phi(1, 0))) = \exp(0) = 1 \). Also there is only one left-linking tree possible (with item 0 as the root and 1 as its only child), and so \( T(1, w, \gamma) = \exp(\frac{1}{\gamma}(w \cdot \phi(1, 0))) = 1 \). Thus the hypothesis holds for \( m_d = 1 \).

Now, lets assume that the induction hypothesis holds for \( m_d = n - 1 \) for \( n \geq 2 \). That is we have

\[ T(n - 1, w, \gamma) = Z(n - 1, w, \gamma) \]

\[ \Rightarrow \sum_{z \in \mathcal{Z}^{n-1}} \exp \left( \frac{1}{\gamma} \sum_{i,j \in z} w \cdot \phi(i, j) \right) \]

\[ = \prod_{i=1}^{n-1} \left( \sum_{\sigma \leq j < i} \exp \left( \frac{1}{\gamma} \sum_{i,j \in z} w \cdot \phi(i, j) \right) \right), \]

where \( \mathcal{Z}^{n-1} \) is the set of left-linking trees over \( n - 1 \) items.

Our goal is to prove the same holds for \( m_d = n \). Consider the expression for \( T(n, w, \gamma) \):

\[ \sum_{z \in \mathcal{Z}^n} \exp \left( \frac{1}{\gamma} \sum_{i,j \in z} w \cdot \phi(i, j) \right), \]

where \( \mathcal{Z}^n \) is the set of left-linking trees over \( n \) items. Notice that for a left-linking tree, the edge connecting item \( n \) to its parent is independent of the remaining edges. In other words, \( z \) is a valid left-linking tree over \( n \) items iff by removing the item \( n \) and its associated edge, we get a valid left-linking tree over \( n - 1 \) items. Thus we can construct \( \mathcal{Z}^n \), the set of all left-linking trees over \( n \) items by taking \( \mathcal{Z}^{n-1} \), the set of all left-linking trees over \( n - 1 \)
items, and connecting item \( n \) to any of the previous \( n \) items \((0, \ldots, n-1)\) i.e. \( Z^n = \{ z \cup \{(n,j)\} | z \in Z^{n-1}, j \in \{0, \ldots, n-1\}\} \). This implies that we can re-write the expression in Eq. (4) as
\[
\sum_{0 \leq j < n, \ z' \in Z^{n-1}} \left( \exp \left( \frac{1}{\gamma} (w \cdot \phi(n,j)) \right) \right)
\times \exp \left( \frac{1}{\gamma} \left( \sum_{i,k \in z'} w \cdot \phi(i,k) \right) \right)
= \sum_{0 \leq j < n} \exp \left( \frac{1}{\gamma} (w \cdot \phi(n,j)) \right)
\times \left( \prod_{z' \in Z^{n-1}} \exp \left( \frac{1}{\gamma} \left( \sum_{i,k \in z'} w \cdot \phi(i,k) \right) \right) \right)
= \prod_{i=1}^{n-1} \left( \sum_{0 \leq k < i} \exp \left( \frac{1}{\gamma} (w \cdot \phi(i,j)) \right) \right),
\]
which is the same as \( Z(n, w, \gamma) \). Hence our proof is complete and the two partition functions are the same.

One can analogously prove that the numerator of Equations (1) and (2) are the same i.e.
\[
Z(C; d, w, \gamma) = \prod_{i=1}^{m} \left( \sum_{0 \leq j < i} \exp \left( \frac{1}{\gamma} (w \cdot \phi(i,j)) \right) C(i,j) \right)
= \sum_{z \in Z_d} \exp \left( \frac{1}{\gamma} \left( \sum_{(i,j) \in z} w \cdot \phi(i,j) \right) \right).
\]
This implies that \( P^{\nu'}[C; d, w] = P^{\nu}[C; d, w] \) for \( \gamma > 0 \).

Now, as \( \gamma \to 0 \), the function in Eq. (1) converge to a Kronecker delta function as explained in the main paper. Also, for \( \gamma \to 0 \), the function in Eq. (2) converges to a Kronecker delta function which is 1 for the clustering consistent with the maximum weight left-linking tree, and 0 else where. As the two probability functions are always bounded and continuous for \( \gamma > 0 \), the equivalence of the two probabilities holds as \( \gamma \to 0 \), where or \( \gamma = 0 \), it is assumed that the functions in Eq. (1) and (2) are replaced by appropriate Kronecker delta functions.

## 2. Results with Randomized Ordering

Table 1 presents results for coreference clustering for ACE data and for document clustering based on authors and topics. Notice that when compared with results in Tables 1 and 2b, the performance declines due to disruption in the natural ordering of items. The deterioration is significant for coreference clustering (approx 3 points decrease in average of MUC, B^3, and CEAF) and for topic-based clustering (approx 10 points increase in VI), but not so significant for author-based clustering (< 1 point increase in VI.)