Supplement: A Discriminative Latent Variable Model for Online Clustering

Rajhans Samdani, Google Research RAJHANS@GOOGLE.COM
Kai-Wei Chang, University of Illinois KCHANG10@ILLINOIS.EDU
Dan Roth, University of Illinois DANR@ILLINOIS.EDU

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 L^3M 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 $\mathcal{C}$ as per the L^3M model as:

$$Pr[\mathcal{C} \lhd i; d, \mathbf{w}] = \sum_{0 \le j < i} Pr[j \leftarrow i; d, \mathbf{w}] \mathcal{C}(i,j) = \frac{Z_i(\mathcal{C}; d, \mathbf{w}, \gamma)}{Z_i(d, \mathbf{w}, \gamma)}, \tag{1}$$

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

For the left-linking tree model, the probability of a left-linking tree $\mathbf{z}$ is represented as

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

where $T(d, w, \gamma) = \sum_{\mathbf{z} \in \mathcal{Z}_d} \exp\left(\frac{1}{\gamma}\left(\sum_{(i,j)\in \mathbf{z}} \mathbf{w} \cdot \phi(i,j)\right)\right)$ is the left-linking tree partition function. The probability of a clustering $\mathcal{C}$ 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 $\mathcal{C}$:

$$\begin{aligned} Pr'[\mathcal{C}; d, \mathbf{w}] &= \sum_{\mathbf{z} \in \mathcal{Z}_d^c} Pr[\mathbf{z}; d, \mathbf{w}] \\ &= \frac{1}{T(d, w, \gamma)} \sum_{\mathbf{z} \in \mathcal{Z}_d^c} \exp\left(\frac{1}{\gamma}\left(\sum_{(i,j)\in \mathbf{z}} \mathbf{w} \cdot \phi(i,j)\right)\right) . \end{aligned} \tag{2}$$

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 probability of clustering for L^3M as expressed in Eq.* (1)*, i.e. $Pr'[\mathcal{C}; d, \mathbf{w}] = Pr[\mathcal{C}; d, \mathbf{w}]$.*

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 partitions functions are the same: $T(d, \mathbf{w}, \gamma) = Z(d, \mathbf{w}, \gamma)$. The proof for the equivalence of the numerators for Eq. (1) and Eq. (2) will be analogous.

Below, we prove $T(d, \mathbf{w}, \gamma) = Z(d, \mathbf{w}, \gamma)$ by induction on the number of items, m_d. We abuse the notation and write $Z(n, \mathbf{w}, \gamma)$ as the partition function for L^3M when considering only the first n items. We use the notation $T(n, \mathbf{w}, \gamma)$ similarly.

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

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

$$\begin{aligned} & T(n-1, \mathbf{w}, \gamma) = Z(n-1, \mathbf{w}, \gamma) \\ \Rightarrow \quad & \sum_{\mathbf{z} \in \mathcal{Z}^{n-1}} \exp\left(\frac{1}{\gamma}\left(\sum_{(i,j)\in \mathbf{z}} \mathbf{w} \cdot \phi(i,j)\right)\right) \\ & = \prod_{i=1}^{n-1} \left(\sum_{0 \le j < i} \exp\left(\frac{1}{\gamma}\left(\mathbf{w} \cdot \phi(i,j)\right)\right)\right) , \end{aligned} \tag{3}$$

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 $T(n, \mathbf{w}, \gamma)$:

$$\sum_{\mathbf{z} \in \mathcal{Z}^n} \exp\left(\frac{1}{\gamma}\left(\sum_{(i,j)\in \mathbf{z}} \mathbf{w} \cdot \phi(i,j)\right)\right) , \tag{4}$$

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, $\mathbf{z}$ 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 $n-1$

Task →	ACE Coreference				Author Clustering	Topic Clustering w/ one pass	
Technique	MUC	B^3	CEAF	AVG	Varation of Information (VI)		
Sum-Link	69.35	77.3	73.54	73.40	134.03	259.57	263.62
Bin-Left-Link	72.62	76.84	74.89	74.78	133.62	252.69	257.69
L^3M ($\gamma = 0$)	75.18	78.66	76.02	76.62	133.66	252.17	254.64
L^3M (tuned γ)	75.47	79.1	76.16	76.91	132.81	245.33	249.09

Table 1: Results with randomized ordering of items on coreference resolution for the ACE data and on author-based and discussion topic-based clustering for Forum data. All the results are scaled by 100. Compare these results to the non-randomized results in the main paper (Tables 1 and 2b in the main paper.) The results for coreference resolution and topic clustering get significantly worse after randomization, while the effect is not so pronounced for author-based clustering.

items, and connecting item n to any of the previous n items $(0, \ldots, n-1)$ i.e. $\mathcal{Z}^n = \{\mathbf{z} \cup \{(n,j)\} | \mathbf{z} \in \mathcal{Z}^{n-1}, j \in \{0, \ldots, n-1\}\}$. This implies that we can re-write the expression in Eq. (4) as

$$\begin{aligned}
& \textstyle\sum_{0\le j<n,\ \mathbf{z}'\in\mathcal{Z}^{n-1}} \left(\exp\left(\frac{1}{\gamma}\left(\mathbf{w}\cdot\phi(n,j)\right)\right) \right. \\
\times\ & \left. \exp\left(\frac{1}{\gamma}\left(\sum_{(i,k)\in\mathbf{z}'} \mathbf{w}\cdot\phi(i,k)\right)\right)\right) \\
=\ & \left(\textstyle\sum_{0\le j<n} \exp\left(\frac{1}{\gamma}\left(\mathbf{w}\cdot\phi(n,j)\right)\right)\right) \\
\times\ & \left(\textstyle\sum_{\mathbf{z}\in\mathcal{Z}^{n-1}} \exp\left(\frac{1}{\gamma}\left(\sum_{(i,k)\in\mathbf{z}'} \mathbf{w}\cdot\phi(i,k)\right)\right)\right) \\
=\ & \left(\textstyle\sum_{0\le j<n} \exp\left(\frac{1}{\gamma}\left(\mathbf{w}\cdot\phi(n,j)\right)\right)\right) \\
\times\ & \left(\prod_{i=1}^{n-1}\left(\textstyle\sum_{0\le k<i} \exp\left(\frac{1}{\gamma}\left(\mathbf{w}\cdot\phi(i,k)\right)\right)\right)\right) \quad \text{(by Eq. (3))} \\
=\ & \prod_{i=1}^{n}\left(\textstyle\sum_{0\le j<i} \exp\left(\frac{1}{\gamma}\left(\mathbf{w}\cdot\phi(i,j)\right)\right)\right),
\end{aligned}$$

which is the same as $Z(n, \mathbf{w}, \gamma)$. 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.

$$\begin{aligned}
Z(\mathcal{C}; d, \mathbf{w}, \gamma) &= \prod_{i=1}^{m_d}\left(\sum_{0\le j<i} \exp\left(\frac{1}{\gamma}\left(\mathbf{w}\cdot\phi(i,j)\right)\right)\mathcal{C}(i,j)\right) \\
&= \sum_{\mathbf{z}\in\mathcal{Z}_d^c} \exp\left(\frac{1}{\gamma}\left(\sum_{(i,j)\in\mathbf{z}} \mathbf{w}\cdot\phi(i,j)\right)\right).
\end{aligned}$$

This implies that $Pr'[\mathcal{C}; d, \mathbf{w}] = Pr[\mathcal{C}; d, \mathbf{w}]$ 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.)