Supplementary Material: Lexical and Hierarchical Topic Regression

Viet-An Nguyen
Computer Science
University of Maryland
College Park, MD
vietan@cs.umd.edu

Jordan Boyd-Graber
iSchool & UMIACS
University of Maryland
College Park, MD
jbg@umiacs.umd.edu

Philip Resnik
Linguistics & UMIACS
University of Maryland
College Park, MD
resnik@umd.edu

D	# documents
S_d	# sentences in document d
$S_{d,t}$	# groups (i.e. sentences) sitting at table t in restaurant d
$N_{d,s}$	# tokens $\boldsymbol{w}_{d,s}$
$N_{d,\cdot,l}$	# tokens in $\boldsymbol{w}_d$ assigned to level l
$N_{d,\cdot,>l}$	# tokens in $\boldsymbol{w}_d$ assigned to level $> l$
$N_{d,\cdot,\geq l}$	$\equiv N_{d,\cdot,l} + N_{d,\cdot,>l}$
$M_{c,l}$	# tables at level l on path c
$C_{c,l,v}$	# word type v assigned to level l on path c
$C_{d,x,l,v}$	# word type v in $\boldsymbol{v}_{d,x}$ assigned to level l
ϕ_k	Topic at node k
η_k	Regression parameter at node k
τ_v	Regression parameter of word type v
$c_{d,t}$	Path assignment for table t in restaurant d
$t_{d,s}$	Table assignment for group $\boldsymbol{w}_{d,s}$
$z_{d,s,n}$	Level assignment for $w_{d,s,n}$
$k_{d,s,n}$	Node assignment for $w_{d,s,n}$ (i.e., node at level $z_{d,s,n}$ on path $c_{d,t_{d,s}}$)
L	Height of the tree
$\mathcal{C}^+$	Set of all possible paths (including new ones) of the tree

Table 1: Notation used for SHLDA's model and inference

This supplementary material provides more detail for the inference algorithm described in the main paper. First, we expand the two probabilities defined in Equations 1 and 2.

Equation 1 defines the conditional density of an arbitrary set of tokens $\boldsymbol{v}_{d,x}$ (e.g., a token, a sentence or a set of sentences) in document d being assigned to path c given all other assignments.

$$\begin{aligned}
f_c^{-d,x}(\boldsymbol{v}_{d,x}) &\equiv P(\boldsymbol{v}_{d,x} \mid \boldsymbol{v}^{-d,x}, \boldsymbol{c}_{d,x}, \boldsymbol{c}^{-d,x}, \boldsymbol{t}, \boldsymbol{z}) \\
&= \prod_{l=1}^{L} P(\boldsymbol{v}_{d,x,l} \mid \boldsymbol{v}^{-d,x,l}, \boldsymbol{c}_{d,x}, \boldsymbol{c}^{-d,x}, \boldsymbol{t}, \boldsymbol{z}) \\
&= \prod_{l=1}^{L} \frac{P(\boldsymbol{v}_{d,x,l}, \boldsymbol{v}^{-d,x,l} \mid \boldsymbol{c}_{d,x}, \boldsymbol{c}^{-d,x}, \boldsymbol{t}, \boldsymbol{z})}{P(\boldsymbol{v}^{-d,x,l} \mid \boldsymbol{c}_{d,x}, \boldsymbol{c}^{-d,x}, \boldsymbol{t}, \boldsymbol{z})} \\
&= \prod_{l=1}^{L} \frac{\int P(\boldsymbol{v}_{d,x,l}, \boldsymbol{v}^{-d,x,l} \mid \phi_{c,l})\, P(\phi_{c,l} \mid \beta_l)\, d\phi_{c,l}}{\int P(\boldsymbol{v}^{-d,x,l} \mid \phi_{c,l})\, P(\phi_{c,l} \mid \beta_l)\, d\phi_{c,l}} \\
&= \prod_{l=1}^{L} \frac{\Gamma(C_{c,l,\cdot}^{-d,x} + V\beta_l)}{\Gamma(C_{c,l,\cdot}^{-d,x} + C_{d,x,l,\cdot} + V\beta_l)} \prod_{v=1}^{V} \frac{\Gamma(C_{c,l,v}^{-d,x} + C_{d,x,l,v} + \beta_l)}{\Gamma(C_{c,l,v}^{-d,x} + \beta_l)}
\end{aligned} \tag{A.1}$$

where we use $\boldsymbol{v}_{d,x,l}$ to denote the set of tokens in $\boldsymbol{v}_{d,x}$ that are assigned to level l. $C_{c,l,v}$ is the number of times word type v is assigned to node at level l on path c. $C_{d,x,l,v}$ is the number of times word type v in $\boldsymbol{v}_{d,x}$ is assigned to node at level l on path c. Superscript $^{-d,x}$ denotes the same count excluding the assignments of $\boldsymbol{v}_{d,x}$. Marginal counts are represented by $\cdot$'s.

Equation 2 defines the conditional density of the response variable y_d of document d given the set of tokens $\boldsymbol{v}_{d,x}$ assigned to path c and all other assignments

$$g_c^{-d,x}(y_d) \equiv P(y_d \,|\, \boldsymbol{c}_{d,x}, \boldsymbol{c}^{-d,x}, \boldsymbol{z}, \boldsymbol{t})$$

$$= \mathcal{N}\left(\frac{1}{N_{d,\cdot}}\Bigg(\underbrace{\sum_{w_{d,s,n}\in\{\boldsymbol{w}_d \setminus \boldsymbol{v}_{d,x}\}} \eta_{c_{d,t_{d,s}},z_{d,s,n}}}_{\text{other words' topic regressions}} + \underbrace{\sum_{l=1}^{L} \eta_{c,l}\cdot C_{d,x,l,\cdot}}_{\boldsymbol{v}_{d,x}\text{'s topic regression}} + \underbrace{\sum_{s=1}^{S_d}\sum_{n=1}^{N_{d,s}} \tau_{w_{d,s,n}}}_{\text{document lexical regressions}}\Bigg), \rho\right) \tag{A.2}$$

For new node at level l on a new path c^{new}, we integrate over all possible values of $\eta_{c^{new},l}$ by using the following property of Gaussian distribution

$$\int \mathcal{N}(a+bx; y, \sigma_x)\,\mathcal{N}(y; \mu, \sigma_y)\,dy = \mathcal{N}(a+bx; \mu, b^2\sigma_x + \sigma_y)$$

Sampling $\boldsymbol{t}$: For each group (i.e., sentence) $\boldsymbol{w}_{d,s}$, we need to sample a table $t_{d,s}$. The conditional distribution of a table t in restaurant d given $\boldsymbol{w}_{d,s}$ and other assignments is

$$P(t_{d,s}=t \,|\, \text{rest}) \propto P(t_{d,s}=t \,|\, \boldsymbol{t}_d^{-s}) \cdot P(\boldsymbol{w}_{d,s}, y_d \,|\, t_{d,s}=t, \boldsymbol{w}^{-d,s}, \boldsymbol{t}^{-d,s}, \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{\eta}) \tag{A.3}$$

The first factor is the prior probability of a table in a Chinese restaurant process. For an existing table, this probability is proportional to the number of groups currently sitting at that table. For a new table, this is proportional to α, as follow

$$P(t_{d,s}=t \,|\, \boldsymbol{t}_d^{-s}) \propto \begin{cases} S_{d,t}^{-d,s}, & \text{for existing table } t; \\ \alpha, & \text{for new table } t^{new}. \end{cases}$$

The second factor is the joint probability of observing $\boldsymbol{w}_{d,s}$ and y_d given $\boldsymbol{w}_{d,s}$ being assigned to table $t_{d,s}=t$. If t is an existing table, this can be easily computed by multiplying Equation A.1 with Equation A.2. For a new table, we need to sum over all possible path $\mathcal{C}^+$ of the tree, including new ones.

$$P(\boldsymbol{w}_{d,s}, y_d \,|\, t_{d,s}=t, \boldsymbol{w}^{-d,s}, \boldsymbol{t}^{-d,s}, \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{\eta})$$
$$\propto \begin{cases} f_{c_{d,t}}^{-d,s}(\boldsymbol{w}_{d,s}) \cdot g_{c_{d,t}}^{-d,s}(y_d), & \text{for existing table } t; \\ \sum_{c\in\mathcal{C}^+} P(c_{d,t^{new}}=c \,|\, \boldsymbol{c}^{-d,s}) \cdot f_c^{-d,s}(\boldsymbol{w}_{d,s}) \cdot g_c^{-d,s}(y_d), & \text{for new table } t^{new}. \end{cases}$$

where $P(c_{d,t^{new}}=c \,|\, \boldsymbol{c}^{-d,s})$ is the prior probability of a path c, which is

$$P(c_{d,t^{new}}=c \,|\, \boldsymbol{c}^{-d,s}) \propto$$
$$\begin{cases} \prod_{l=2}^{L} \dfrac{M_{c,l}^{-d,s}}{M_{c,l-1}+\gamma_{l-1}}, & \text{for an existing path } c; \\ \dfrac{\gamma_{l^*}}{M_{c^{new},l^*}^{-d,s}+\gamma_{l^*}} \prod_{l=2}^{l^*} \dfrac{M_{c^{new},l}^{-d,s}}{M_{c^{new},l-1}^{-d,s}+\gamma_{l-1}}, & \begin{array}{l}\text{for a new path } c^{new} \text{ which consists of an existing path} \\ \text{from the root to a node at level } l^* \text{ and a new node.}\end{array} \end{cases} \tag{A.4}$$

Here we use $M_{c,l}$ to denote the number of tables assigned to node at level l on path c. As usual, the superscript $^{-d,s}$ denotes the same count but excluding assignments of $\boldsymbol{w}_{d,s}$.

Sampling $\boldsymbol{z}$: After assigning a sentence $\boldsymbol{w}_{d,s}$ to a table, we assign each token $w_{d,s,n}$ to a level to choose a dish from the combo associated with the table. The probability of assigning $w_{d,s,n}$ to level l conditioning on other assignments is

$$P(z_{d,s,n}=l \,|\, \text{rest}) \propto P(z_{d,s,n}=l \,|\, \boldsymbol{z}_d^{-s,n}) \cdot P(w_{d,s,n}, y_d \,|\, z_{d,s,n}=l, \boldsymbol{w}^{-d,s,n}, \boldsymbol{z}^{-d,s,n}, \boldsymbol{t}, \boldsymbol{c}, \boldsymbol{\eta}) \tag{A.5}$$

The first factor captures the probability that a customer in restaurant d is assigned to level l, conditioned on the level assignments of all other customers in restaurant d. Since the level distribution is modeled using a truncated stick breaking prior $\text{GEM}(m, \pi)$, this probability is the posterior expected value of the l^{th} weight from the stick [1]

$$P(z_{d,s,n} = l \mid \boldsymbol{z}_d^{-s,n}) = \frac{m\pi + N_{d,\cdot,l}^{-d,s,n}}{\pi + N_{d,\cdot,\geq l}^{-d,s,n}} \prod_{j=1}^{l-1} \frac{(1-m)\pi + N_{d,\cdot,>j}^{-d,s,n}}{\pi + N_{d,\cdot,\geq j}^{-d,s,n}},$$

where $N_{d,\cdot,l}$ is the number of tokens in document d assigned to level l; $N_{d,\cdot,>l}$ is the number of tokens in document d assigned to level $> l$; and $N_{d,\cdot,\geq l} \equiv N_{d,\cdot,l} + N_{d,\cdot,>l}$.

The second factor is the probability of observing $w_{d,s,n}$ and y_d, conditioning on $w_{d,s,n}$ being assigned to level l and other assignments. This is computed using Equations A.1 and A.2 as follow

$$P(w_{d,s,n}, y_d \mid z_{d,s,n} = l, \boldsymbol{w}^{-d,s,n}, \boldsymbol{z}^{-d,s,n}, \boldsymbol{t}, \boldsymbol{c}, \boldsymbol{\eta}) = f_{c_{d,t_{d,s}}}^{-d,s,n}(w_{d,s,n}) \cdot g_{c_{d,t_{d,s}}}^{-d,s,n}(y_d).$$

Sampling $\boldsymbol{c}$: After assigning customers to tables and levels, we also sample the path assignments for all tables. This is important since it potentially changes the assignments of all customers sitting at a given table, which leads to a well-mixed Markov chain and faster convergence. The probability of assigning a table t in restaurant d to a path c is

$$P(c_{d,t} = c \mid \text{rest}) \propto P(c_{d,t} = c \mid \boldsymbol{c}^{-d,t}) \cdot P(\boldsymbol{w}_{d,t}, y_d \mid c_{d,t} = c, \boldsymbol{w}^{-d,t}, \boldsymbol{c}^{-d,t}, \boldsymbol{t}, \boldsymbol{z}, \boldsymbol{\eta}) \qquad \text{(A.6)}$$

where we slightly abuse the notation by using $\boldsymbol{w}_{d,t} \equiv \cup_{\{s \mid t_{d,s}=t\}} \boldsymbol{w}_{d,s}$ to denote the set of customers in all the groups sitting at table t in restaurant d. The first factor is the prior probability of a path given all tables' path assignments $\boldsymbol{c}^{-d,t}$, excluding table t in restaurant d and is computed using Equation A.4

The second factor in Equation A.6 is the probability of observing $\boldsymbol{w}_{d,t}$ and y_d given the new path assignments, $P(\boldsymbol{w}_{d,t}, y_d \mid c_{d,t} = c, \boldsymbol{w}^{-d,t}, \boldsymbol{c}^{-d,t}, \boldsymbol{t}, \boldsymbol{z}, \boldsymbol{\eta}) = f_c^{-d,t}(\boldsymbol{w}_{d,t}) \cdot g_c^{-d,t}(y_d)$.

Optimizing $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$: We optimize the regression parameters $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ via the likelihood

$$\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\tau}) = -\frac{1}{2\rho} \sum_{d=1}^{D} (y_d - \boldsymbol{\eta}^T \bar{\boldsymbol{z}}_d - \boldsymbol{\tau}^T \bar{\boldsymbol{w}}_d)^2 - \frac{1}{2\sigma} \sum_{k=1}^{K^+} (\eta_k - \mu)^2 - \frac{1}{\omega} \sum_{v=1}^{V} |\tau_v|, \qquad \text{(A.7)}$$

The derivatives of this objective function with respect to each η_k is

$$\frac{\mathrm{d}\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\tau})}{\mathrm{d}\eta_k} = -\frac{1}{\rho} \sum_{d=1}^{D} \bar{z}_{d,k} \cdot (\boldsymbol{\eta}^T \bar{\boldsymbol{z}}_d + \boldsymbol{\tau}^T \bar{\boldsymbol{w}}_d - y_d) - \frac{1}{\sigma}(\eta_k - \mu)$$

Since the L1-norm on $\boldsymbol{\tau}$ makes $\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\tau})$ non-differentiable when $\tau_v = 0$, we use the sub-gradient strategy [2] to approximate the gradient. Another heuristic to address this problem is to manually set the value of τ_v equal 0 for a subset of word types v in the vocabulary that have high TF-IDFs and perform L2-norm regularization on the remaining word types. We found that using this heuristic works reasonably well in practice and is able to speed up the optimization procedure significantly (depending on how large the set of word types that have $\tau_v = 0$).

References

[1] Porteous, I., A. Ihter, P. Smyth, et al. Gibbs sampling for (coupled) infinite mixture models in the stick breaking representation. In *UAI*. 2006.

[2] Schmidt, M., G. Fung, R. Rosales. Fast optimization methods for L1 regularization: A comparative study and two new approaches. In *ECML*. 2007.