Appendix to Author Clustering and Topic Estimation on Short Texts

A Collapsed Gibbs Sampler

Proof for the collapsed Gibbs sampler. Z_{ud} is the topic of tweet d by user u. θ_u is the user's topic distribution, β_t is the topic distribution t over words. We desire $P(Z_{ud} = t|\mathbf{Z_{-ud}}, \mathbf{W})$, the distribution of Z_{ud} given the topics of all tweets except tweet d by user u and the word counts of all tweets $\mathbf{W}$. Given the topic of a tweet $Z_{ud} = t$, the word counts $\mathbf{W_{ud}}$ are a multinomial draw from topic distribution β_t. θ_u identifies the unconditional, prior distribution for Z_{ud}. Via Bayes' rule, the full conditional for $P(Z_{ud} = t|\theta_u, \beta_t, \mathbf{Z_{-ud}}, \mathbf{W})$ proportional to the product of this multinomial likelihood and θ_u prior. Thus, given β_t and θ_u, the posterior for Z_{ud} does not depend on any words or topics besides those in tweet ud.

$$P(Z_{ud} = t|\theta_u, \beta_t, \mathbf{Z_{-ud}}, \mathbf{W}) \propto P(\mathbf{W_{ud}}|Z_{ud} = t, \beta_t)P(Z_{ud} = t|\theta_t) \propto \prod_i \beta_{ti}^{W_{udi}} \theta_{ut}$$

From the paper:

$$\begin{aligned}
&p(Z_{ud} = t|\mathbf{Z_{-ud}}, \mathbf{W}) \\
&\quad = \int\int p(Z_{ud} = t|\theta_u, \beta_t, \mathbf{W_{ud}})p(\theta_u, \beta_t|\mathbf{Z_{-ud}}, \mathbf{W_{-ud}})\ d\theta_u d\beta_t \\
&\quad \propto \int\int \theta_{ut} \prod_i \beta_{ti}^{W_{udi}} p(\theta_t|\mathbf{Z_{-ud}}) \times \\
&\qquad\qquad p(\beta_t|\mathbf{Z_{-ud}}, \mathbf{W_{-ud}})\ d\theta_u d\beta_t \\
&\quad = \int \theta_{ut} p(\theta_u|\mathbf{Z_{-ud}})\ d\theta_u \times \\
&\qquad \int \prod_i \beta_{ti}^{W_{udi}} p(\beta_t|\mathbf{Z_{-ud}}, \mathbf{W_{-ud}})\ d\beta_t
\end{aligned}$$

Conditional on $G_u = g$ and α_g, the unconditional distribution of θ_u is: $p(\theta_u) \sim \text{Dir}(\alpha_g)$. The Dirichlet is conjugate with multinomial data. Let $\mathbf{Z_{-d}^{(u)}}$ denote the topic counts of user u excluding the topic of tweet d.

$$\begin{aligned}
p(\theta_u|G_u = g, \alpha_g, \mathbf{Z_{-ud}}) &\propto p(\mathbf{Z^{(u)}_{-d}}|\theta_u)p(\theta_u|G_u = g, \alpha_g) \\
&\propto \prod_{t=1}^{T} \theta_{ut}^{Z^{(u)}_{-d,t}} \prod_{t=1}^{T} \theta_{ut}^{\alpha_{gt}-1} \\
&= \prod_{t=1}^{T} \theta_{ut}^{Z^{(u)}_{-d,t}+\alpha_{gt}-1} \\
&\propto \mathrm{Dir}(\mathbf{Z^{(u)}_{-d}} + \alpha_g)
\end{aligned}$$

Using this result in the first integral, we have:

$$\begin{aligned}
\int \theta_{ut} p(\theta_u|\mathbf{Z_{-ud}})\ d\theta_u &= \frac{\Gamma\left(\sum_j Z^{(u)}_{-d,j} + \alpha_{gj}\right)}{\prod_j \Gamma\left(Z^{(u)}_{-d,j} + \alpha_{gj}\right)} \int \theta_{ut} \prod_{j=1}^{T} \theta_{uj}^{Z^{(u)}_{-d,j}+\alpha_{gj}-1}\ d\theta_u \\
&= \frac{\Gamma\left(\sum_j Z^{(u)}_{-d,j} + \alpha_{gj}\right)}{\prod_j \Gamma\left(Z^{(u)}_{-d,j} + \alpha_{gj}\right)} \frac{\prod_j \Gamma\left(Z^{(u)}_{-d,j} + \alpha_{gj} + I(j = t)\right)}{\Gamma\left(\sum_t Z^{(u)}_{-d,j} + \alpha_{gj} + 1\right)} \\
&= \frac{Z^{(u)}_{-d,t} + \alpha_{gt}}{\sum_j Z^{(u)}_{-d,j} + \alpha_{gj}}
\end{aligned}$$

This is the result in the paper. $\Gamma(\cdot)$ denotes the gamma function and $I(\cdot)$ the indicator function. The integral is computed by recognizing the kernel of a Dirichlet distribution with the same parameters as the posterior with 1 added the t'th parameter.

The second integral is computed similarly. For documents with $Z_{ud} = t$, the word counts are a multinomial draw of size n from the probability distribution β_t. Thus, the posterior can be expressed as: $p(\beta_t|\mathbf{Z_{-ud}}, \mathbf{W_{-ud}}) \propto p(\beta_t|\mathbf{W^{(t)}_{-ud}})$, conditioning only on the word counts from documents with $Z_{ud} = t$. With a $\mathrm{Dir}(\eta)$ prior on β_t, the posterior also Dirichlet with parameters $\mathbf{W^{(t)}_{-ud}} + \eta$ by the same conjugate update shown above. Thus, the second integral simplifies to:

$$\begin{aligned}
\int \prod_i \beta_{ti}^{W_{ud,i}} p(\beta_t|\mathbf{Z_{-ud}}, \mathbf{W_{-ud}})\ d\beta_t &= \frac{\Gamma\left(N\eta + \sum_i^N W^{(t)}_{-ud,i}\right)}{\prod_i^N \Gamma\left(\eta + W^{(t)}_{-ud,i}\right)} \int \prod_i \beta_{ti}^{W_{ud,i}} \prod_i \beta_{ti}^{W_{-ud,i}+\eta}\ d\beta_t \\
&= \frac{\Gamma\left(N\eta + \sum_i^N W^{(t)}_{-ud,i}\right)}{\prod_i^N \Gamma\left(\eta + W^{(t)}_{-ud,i}\right)} \frac{\prod_i^N \Gamma\left(\eta + W^{(t)}_{-ud,i} + W_{ud,i}\right)}{\Gamma\left(N\eta + \sum_i^N W^{(t)}_{-ud,i} + W_{ud,i}\right)}
\end{aligned}$$

This is the result in the paper. The integral is computed by recognizing the kernel of a Dirichlet distribution with the same parameters as the posterior for β_t with the word counts for tweet ud

added to the respective parameters. Some simplification of the gamma functions are possible. They do not add to the intuitive understanding, so we do not show them here.

B Variational stLDA-C

This section derives the variational approximation to the posterior from Section 3.2 in the paper. This follows the same general procedure as the variational approximation derived in Blei (2003).

B.1 ELBO Bound

Here we derive a evidence lower bound (ELBO) on the KL divergence between the variational distribution and the posterior. First, we will build the ELBO for an individual user after re-introducing the notation from the main paper.

$\mathbf{W}_u$ is the $(n_u \times V)$ document-term matrix for user u. The latent variables are: g_u their cluster membership, θ_u their topic distribution, and $\mathbf{Z}_u$ the topics of their posts. For simplicity, I omit the $\_u$ notation for the remainder of this section. The model parameters are: α the $G \times T$ matrix of Dirichlet parameters for each cluster, β the $T \times V$ matrix of topic distributions, and ξ the corpus-level cluster proportions. The variational distribution q is expressed as:

$$q(g, \theta, \mathbf{Z}|\lambda, \gamma, \phi) = q(g|\lambda)q(\theta|\gamma)\prod_{d=1}^{n_u} q(z_d|\phi_d)$$

Here $q(g|\lambda)$ is $\text{Cat}(\lambda)$, $q(\theta|\gamma)$ is $\text{Dir}(\gamma)$, and $q(z_d|\phi_d)$ is $\text{Cat}(\phi_d)$. Both λ and ϕ_d are points on the probability simplex; γ is a T dimensional vector of positive numbers. Although not made explicit in the notation, $\{\lambda, \gamma, \phi\}$ are all *user-specific* quantities. For consistency with Blei (2003) notation, I let ξ denote the cluster mixing proportions (ϕ in the main paper) and let ϕ_d be the post-specific variational parameters. Now, we express the probability of the user's posts.

$$\begin{aligned}
\log p(\mathbf{W}|\alpha, \beta, \xi) &= \log \sum_g \int_\theta \sum_z p(\theta, \mathbf{Z}, g|\alpha, \beta, \xi) \\
&= \log \sum_g \int_\theta \sum_z p(\theta, \mathbf{Z}, g|\alpha, \beta, \xi) \frac{q(g, \theta, \mathbf{Z})}{q(g, \theta, \mathbf{Z})} \\
&= \log E_q \left[\frac{p(\theta, \mathbf{Z}, g|\alpha, \beta, \xi)}{q(g, \theta, \mathbf{Z})} \right] \\
&\geq E_q[\log p(g, \theta, \mathbf{Z}, \mathbf{W}|\alpha, \beta, \xi)] - E_q[\log q(g, \theta, \mathbf{Z})] := L
\end{aligned}$$

The last step is from applying Jensen's inequality to push the log through the expectation. E_q notes the expectation is taken with respect to the variational distribution q.

As in normal LDA, the difference between the left and right hand terms ($\log p(\mathbf{W}|\alpha,\beta,\xi)$ and the lower bound, L) is the KL divergence between the variational posterior and the true posterior. Let $D(A||B)$ denote the KL divergence between A and B.

$$\begin{aligned} E_q[\log p(g,\theta,\mathbf{Z},\mathbf{W}|\alpha,\beta,\xi)] - E_q[\log q(g,\theta,\mathbf{Z})] &= \log p(\mathbf{W}|\alpha,\beta,\xi) + E_q[\log p(g,\theta,\mathbf{Z}|\mathbf{W},\alpha,\beta,\xi) - E_q[\log q(g,\theta,\mathbf{Z})] \\ &= \log p(\mathbf{W}|\alpha,\beta,\xi) - D(q(g,\theta,\mathbf{Z})||p(g,\theta,\mathbf{Z}|\mathbf{W},\alpha,\beta,\xi)) \end{aligned}$$

The likelihood of the data and latent parameters can be factored into simpler conditional distributions.

$$p(g,\theta,\mathbf{Z},\mathbf{W}|\alpha,\beta,\xi) = p(g|\xi)p(\theta|g,\alpha)p(\mathbf{Z}|\theta)p(\mathbf{W}|\mathbf{Z},\beta)$$

Thus, the ELBO (L) can be expressed as (reintroducing variational parameters):

$$\begin{aligned} L(\lambda,\gamma,\phi;\alpha,\beta,\xi) &= E_q[\log p(g,\theta,\mathbf{Z},\mathbf{W}|\alpha,\beta,\xi)] - E_q[\log q(g,\theta,\mathbf{Z})] \\ &= E_q[\log p(g|\xi)] + E_q[\log p(\theta|g,\alpha)] + E_q[\log p(\mathbf{Z}|\theta)] + E_q[\log p(\mathbf{W}|\mathbf{Z},\beta)] \\ &- E_q[\log q(g)] - E_q[\log q(\theta)] - E_q[\log q(\mathbf{Z})] \end{aligned}$$

Each expectation can be simplified. Each line below corresponds to expansion of the seven terms above, in order.

$$\begin{aligned} L(\lambda,\gamma,\phi;\alpha,\beta,\xi) &= \sum_g \lambda_g \log(\xi_g) && (1) \\ &+ \sum_g \lambda_g \left(\log\Gamma(\sum_{t=1}^T \alpha_{gt}) - \sum_{t=1}^T \log\Gamma(\alpha_{gt}) + \sum_{t=1}^T (\alpha_{gt}-1)E_q[\log\theta_t] \right) && (2) \\ &+ \sum_{d=1}^n \sum_{t=1}^T \phi_{dt} E_q[\log\theta_t] && (3) \\ &+ \sum_{d=1}^n \sum_{t=1}^T \phi_{dt} \left[\log\left(\frac{\left(\sum_{j=1}^V W_{dj}\right)!}{\prod_{j=1}^V W_{dj}!} \right) + \sum_{j=1}^V W_{dj}\log\beta_{tj} \right] && (4) \\ &- \sum_g \lambda_g \log\lambda_g && (5) \\ &- \log\Gamma(\sum_{t=1}^T \gamma_t) + \sum_{t=1}^T \log\Gamma(\gamma_t) - \sum_{t=1}^T (\gamma_t - 1)E_q[\log\theta_t] && (6) \\ &- \sum_{d=1}^n \sum_{t=1}^T \phi_{dt}\log(\phi_{dt}) && (7) \end{aligned}$$

These are nearly the same terms derived by Blei et al. (2003) Appendix A.3, Equation (15). Lines (3), (6), and (7) are the same. Lines (1) and (5) are new for cluster membership. (2) is more complicated because of the summation over the clusters; the $\sum_g$ component is new. (4) is only different because the normalizing constant for the multinomial distribution is added; the post is a draw of size n_d from the relevant topic distribution, rather than each word being a draw of size one. Line (4) is the only change required if stLDA (no clustering) is used rather than traditional LDA.

B.2 Variational Parameter Estimation (λ, γ, ϕ)

Here I derive the ELBO-maximizing values of each variational parameter. Note that, as with vanilla LDA, the update rules are dependent, so the estimation needs to iterate between all of them until convergence.

B.2.1 λ

Recall that λ_g is the variational probability that the user is in cluster g. Maximizing L with respect to λ only involves lines (1), (2), and (5) with the constraint that $\sum_g \lambda_g = 1$. The derivative of L with respect to λ_g is:

$$\log(\xi_g) + f(\alpha_g, \gamma) - \log \lambda_g - 1 + c$$

Where $f(\alpha_g, \gamma)$ is from line (2) above and c is a constant from the Lagrangian. Setting equation equal to zero gives

$$\lambda_g \propto \xi_g \exp(f(\alpha_g, \gamma)) = \xi_g \exp(E_q[\log p(\theta | g, \alpha_g)])$$

This is interpretable as proportional to the "prior" ξ_g times the exponential of the variational expected log likelihood of θ if $\theta \sim Dir(\alpha_g)$. This is interpretable as computing $p(g|\theta)$ as $\propto p(\theta|g)p(g)$ with the likelihood approximated via the variational posterior.

B.2.2 ϕ

Recall that ϕ_{dt} is the variational probability that document d has topic t. Maximizing L with respect to ϕ only involves lines (3), (4), and (7) with the constraint that $\forall d$, $\sum_t \phi_{dt} = 1$. The terms in L that are functions of ϕ_{dt} are isolated below, with c_d to represent the multinomial normalizing constant.

$$\phi_{dt} E_q[\log \theta_t] + \phi_{dt} \log(c_d) + \sum_{j=1}^{V} W_{dj} \log \beta_{tj} - \phi_{dt} \log(\phi_{dt})$$

The derivative of L with respect to ϕ_{dt} including Lagragian term ζ is:

$$E_q[\log \theta_t] + \log(c_d) + \sum_{j=1}^{V} W_{dj} \log \beta_{tj} - \log(\phi_{dt}) - 1 + \zeta$$

Setting equal to zero gives:

$$\log\phi_{dt} = E_q[\log\theta_t] + \log(c_d) + \log\left(\prod_{j=1}^{V}\beta_{tj}^{W_{dj}}\right) - 1 + \zeta \implies$$
$$\phi_{dt} \propto \exp(E_q[\log\theta_t])\prod_{j=1}^{V}\beta_{tj}^{W_{dj}}$$

This is again interpretable as the likelihood times the prior, but this time the prior is approximated via the variational distribution.

B.2.3 γ

Maximizing L with respect to γ involves lines (2), (3), and (6), noting that $E_q[\log\theta_t] = g_t(\gamma)$, a known formula with well established numerical approximation methods. The components of L involving γ_t are listed below:

$$\sum_g \lambda_g \sum_t (\alpha_{gt}-1)g_t(\lambda) + \sum_d\sum_t \phi_{dt}g_t(\lambda) - \log\Gamma\left(\sum_{t=1}^{T}\gamma_t\right) + \log\Gamma(\gamma_t) - \sum_{t=1}^{T}(\gamma_t-1)g_t(\lambda)$$

This simplifies to:[1]

$$\sum_t g_t(\lambda)\left(\sum_g \lambda_g\alpha_{gt} + \sum_d \phi_{dt} - \gamma_t\right) - \log\Gamma\left(\sum_t \gamma_t\right) + \log\Gamma(\gamma_t)$$

Before taking the derivative, it is necessary to expand g_t. $g_t(x) = \Psi(x_t) - \Psi(\sum_t x_t)$ where Ψ is the digamma function (first derivative of the log gamma function). Note that this is the same expression as in Blei (2003) A.3.2 with $\sum_g \lambda_g\alpha_{gt} := \alpha_i$, the sum of each cluster's relevant Dirichlet parameter weighted by (variational) probability of cluster membership λ_g. This will result in the same maximizing value with the substitution because the substitution just changes a constant term. For clarity, however, I will continue to fully derive the result here. For ease of notation, let $\widetilde{\alpha}_t = \sum_g \lambda_g\alpha_{gt}$.

Substituting the expression for g_t and $\widetilde{\alpha}_t$ gives:

$$\sum_t\left(\widetilde{\alpha}_t + \sum_d \phi_{dt} - \gamma_t\right)\left(\Psi(\lambda_t) - \Psi\left(\sum_t \gamma_t\right)\right) - \log\Gamma\left(\sum_t \gamma_t\right) + \log\Gamma(\gamma_t)$$

Taking a derivative with respect to γ_t gives:

[1] The -1 parenthetical terms cancel because $\sum_g \lambda_g = 1$

$$
\begin{aligned}
\frac{\partial L}{\partial \gamma_t} &= \left(\widetilde{\alpha}_t + \sum_d \phi_{dt} - \gamma_t\right) \Psi'(\gamma_t) - \Psi(\gamma_t) \\
&\quad - \sum_i \left(\widetilde{\alpha}_i + \sum_d \phi_{di} - \gamma_i\right) \Psi'\left(\sum_i \gamma_i\right) + \Psi'\left(\sum_i \gamma_i\right) \\
&\quad - \Psi\left(\sum_t \gamma_t\right) + \Psi(\gamma_t) \\
&= \left(\widetilde{\alpha}_t + \sum_d \phi_{dt} - \gamma_t\right) \Psi'(\gamma_t) - \sum_i \left(\widetilde{\alpha}_i + \sum_d \phi_{di} - \gamma_i\right) \Psi'\left(\sum_i \gamma_i\right)
\end{aligned}
$$

Setting equal to zero yields a maximum (for all t) of $\gamma_t = \sum_g \lambda_g \alpha_{gt} + \sum_d \phi_{dt}$.

B.3 Model Parameter Estimation (α, β, ξ)

The above section was just for a single user (the analog of a single document in traditional LDA). To estimate the parameters shared across users, cluster-specific Dirichlet parameters α_g, topic distributions β, and cluster proportions ξ, we have to sum the ELBO bound derived above across users. In another abuse of notation return the $\_u$ do denote individual users.

Let the full likelihood be expressed as:

$$L(\alpha, \beta, \xi) = \sum_u L_u(\lambda_u, \gamma_u, \phi_u; \alpha, \beta, \xi)$$

L_u is the function defined by equations $(1)-(7)$ above with the user-dependence made explicit. This is the function that needs to be maximized with respect to $\{\alpha, \beta, \xi\}$ holding $\{\lambda_u, \gamma_u, \phi_u\}$ constant.

B.3.1 ξ

This is the easiest parameter. It appears only in term (1). Taking a derivative and using a Lagrangian to enforce $\sum_g \xi_g = 1$ gives:

$$\xi_g \propto \sum_u \lambda_{ug}$$

B.3.2 β

This is similar to the variational multinomial in A.4.1 of Blei (2003), with only the added complexity of having multiple words drawn from the same topic.

The full likelihood terms with β plus Lagrangian term is expressed below. $d \in \mathcal{D}_u$ indexes over all documents produced by user u.

$$\sum_u \sum_{d\in\mathcal{D}_u} \sum_t \sum_j \phi_{udt} W_{udj} \log \beta_{tj} - \sum_t \zeta_t \left(1 - \sum_j^V \beta_{tj}\right)$$

Taking derivatives with respect to β_{tj} results in:

$$\beta_{tj} \propto \sum_u \sum_{d\in\mathcal{D}_u} \phi_{udt} W_{udj}$$

The intuition is that β_t is proportional to the word occurrences weighted by the variational probability that the word was drawn from topic t.

B.3.3 α

These terms are more complicated but can be computed in parallel over clusters. Recall that α is a $G \times T$ matrix of positive values. The ELBO terms involving α are in line (2).

$$\sum_u \sum_g \lambda_{ug} \left(\log \Gamma(\sum_{t=1}^T \alpha_{gt}) - \sum_{t=1}^T \log \Gamma(\alpha_{gt}) + \sum_{t=1}^T (\alpha_{gt} - 1) g_t(\gamma_u) \right)$$

Let Ψ denote the digamma function, the first derivative of the log gamma function. Taking a derivative with respect to α_{gt} gives:

$$\frac{\partial L}{\partial \alpha_{gt}} = \sum_u \lambda_{ug} \left[\Psi\left(\sum_t \alpha_{gt}\right) - \Psi(\alpha_{gt}) + g_t(\gamma_u) \right] = M_g \left[\Psi\left(\sum_t \alpha_{gt}\right) - \Psi(\alpha_{gt}) \right] + \sum_u \lambda_{ug} g_t(\gamma_u)$$

Where $M_g = \sum_u \lambda_{ug}$. Note that this depends on $\alpha_{gt'}$ for $t' \neq t$ but not on any $\alpha_{g't'}$ where $g' \neq g$. Thus, the update step can be computed simultaneously for each cluster.

This has the same form as the derivative in traditional LDA so the same Newton-Raphson algorithm can be used based on the structure of the Hessian matrix.

$$\frac{\partial L}{\partial \alpha_{gt} \alpha_{gt'}} = -I(t = t') M_g \Psi'(\alpha_{gt}) + M_g \Psi' \left(\sum_t \alpha_{gt}\right)$$

Thus, the Hessian for cluster g can be expressed as: $\mathbf{H}_g = diag(-M_g \Psi'(\boldsymbol{\alpha}_g)) + \mathbf{1}\mathbf{1}^T M_g \Psi'(\sum_t \alpha_{gt})$, a diagonal matrix plus a constant matrix.