Syntactic Topic Models Supplement

Jordan Boyd-Graber
Department of Computer Science
35 Olden Street
Princeton University
Princeton, NJ 08540
jbg@cs.princeton.edu

David Blei
Department of Computer Science
35 Olden Street
Princeton University
Princeton, NJ 08540
blei@cs.princeton.edu

In this document, we derive mean field variational inference updates for the syntactic topic model. After recapitulating the model, we expand the expectations for each of the terms and then derive updates for each of the variational parameters that maximize the likelihood bound.

1 Model

As a reference, we first summarize the primary variables in our model.

$\pi_{p,i}$ How much we should want to go into topic i given that the parent topic was p. It is parameterized in the variational distribution by ν.

$\theta_{d,i}$ How much we should want to go into topic i given that the document is d. It is parameterized in the variational distribution by γ in the variational distribution.

z_n The topic of the n^{th} word. It is parameterized by ϕ_n in the variational distribution.

β Top level proportions from DP. In the variational distribution, $q(\beta|\beta^*)$ only has support on β^*, which is zero for all indices greater than K.

$\tau_{i,j}$ How much we should want to generate word j given that the word's topic is i.

These variables are generated through the following process:

1. Choose global topic weights $\beta \sim \text{GEM}(\alpha)$
2. For each topic index $k = \{1, \dots\}$:
   (a) Choose topic transition distribution $\pi_k \sim \text{DP}(\alpha_T, \beta)$.
   (b) Choose topic $\tau_k \sim \text{Dir}(\sigma)$
3. For each document $d = \{1, \dots M\}$:
   (a) Choose topic weights $\theta_d \sim \text{DP}(\alpha_D, \beta)$.
   (b) For each sentence in the document:
      i. Choose topic assignment $z_0 \propto \theta_d \pi_{start}$
      ii. Choose root word $w_0 \sim \text{mult}(1, \tau_{z_0})$
      iii. For each additional word w_n and parent p_n, $n \in \{1, \dots d_n\}$
      - Choose topic assignment $z_n \propto \theta_d \pi_{z_{p(n)}}$
      - Choose word $w_n \sim \text{mult}(1, \tau_{z_0})$

2 Variational Distribution

Our variational distribution is factored into

$$q(\beta, z, \theta, \pi, \tau|\beta^*, \phi, \gamma, \nu) = q(\beta|\beta^*) \prod_d q(\theta_d|\gamma_d) \prod_z q(\pi_z|\nu) \prod_n q(z_n|\phi_n), \tag{1}$$

where $q(\beta|\beta^*)$ is not a full distribution but a degenerate point estimate, γ_d and ν_z are variational Dirichlet distributions, and ϕ_n is a topic multinomial for the n^{th} word. Our lower bound on the likelihood is $L(\gamma, \nu, \phi; \beta, \theta, \pi, \tau)$

$$\mathbb{E}_q\left[\log p(\beta|\alpha) + \log p(\boldsymbol{\theta}|\alpha_D, \beta) + \log p(\boldsymbol{\pi}|\alpha_P, \beta) + \log p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\pi}) + \right.$$
$$\left. \log p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\tau}) + \log p(\boldsymbol{\tau}|\sigma)\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\theta}) + \log q(\boldsymbol{\pi}) + \log q(\boldsymbol{z})\right]. \quad (2)$$

3 Document-specific terms

In our implementation, document-specific variational parameters and global variational parameters are computed separately because the document-specific parameters can be computed in parallel, greatly speeding the inference process. We present the derivation of the variational updates in a similar manner, focusing on document specific terms before global ones because it more closely follows the flow of the implemented algorithm and removes unneeded subscripts and summations.

3.1 Expanding expectations

We first expand the expectation of the per-word topic term, both because it is the most difficult and because it is the crux of this work. Rather than drawing the topic of a word directly from a multinomial, it is chosen from the renormalized point-wise product of two multinomial distributions. In order to handle the expectation of the log sum introduced by the renormalization, we introduce an additional variational parameter ω_n for each word via a Taylor approximation of the logarithm to find that

$$\begin{aligned}
\mathbb{E}_q\left[\log p(\mathbf{z}|\theta, \pi)\right] &= \mathbb{E}_q\left[\log \prod_{n=1}^{N} \frac{\theta_{z_n}\pi_{z_{p(n)},z_n}}{\sum_i^K \theta_i \pi_{z_{p(n)},i}}\right] \\
&= \mathbb{E}_q\left[\sum_{n=1}^{N} \log \theta_{z_n}\pi_{z_{p(n)},z_n} - \sum_{n=1}^{N} \log \sum_{i=1}^{K} \theta_i \pi_{z_{p(n)},i}\right] \\
&\geq \sum_{n=1}^{N} \mathbb{E}_q\left[\log \theta_{z_n}\pi_{z_{p(n)},z_n}\right] - \sum_{n=1}^{N} \mathbb{E}_q\left[\omega_n^{-1} \sum_{i=1}^{K} \theta_i \pi_{z_{p(n)},i}\right] + \log \omega_n - 1 \\
&= \sum_{n=1}^{N} \mathbb{E}_q\left[\log \theta_{z_n}\right] + \mathbb{E}_q\left[\log \pi_{z_{p(n)},z_n}\right] - \left(\sum_{n}^{N} \omega_n^{-1} \sum_{i}^{K} \mathbb{E}_q\left[\theta_i \pi_{z_{p(n)},z_n}\right] + \log \omega_n - 1\right) \\
&= \sum_{n=1}^{N}\sum_{i=1}^{K} \phi_{n,i}\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)\right) + \sum_{n=1}^{N}\sum_{i=1}^{K}\sum_{j=1}^{K} \phi_{n,i}\phi_{p(n),j}\left(\Psi(\nu_{j,i}) - \Psi(\sum_{k=1}^{K}\nu_{j,k})\right) \\
&\quad - \left(\sum_{n=1}^{N} \omega_n^{-1} \sum_{i=1}^{K}\sum_{j=1}^{K} \phi_{p(n),j} \frac{\gamma_i \nu_{j,i}}{\sum_{k=1}^{K}\gamma_k \sum_{k=1}^{K}\nu_{j,k}} + \log \omega_n - 1\right). \quad (3)
\end{aligned}$$

Because the words come from a multinomial distribution, the probability of a word coming from a topic is simply its weight under the multinomial. So taking the expectation over its possible assignments, we have

$$\mathbb{E}_q\left[\log p(\mathbf{w}|\mathbf{z}, \tau, \rho, t)\right] = \mathbb{E}_q\left[\sum_{n=1}^{N} \log \tau_{z,w_n}\right] = \sum_{n=1}^{N}\sum_{i=1}^{K} \phi_i \log \tau_{i,w_n}. \quad (4)$$

Finally, we take the expectation of the document topic multinomial; this is easier because the variational distribution only has support at β^*; this gives us

$$\begin{aligned}
\mathbb{E}_q\left[p(\theta_d|\alpha_D, \beta)\right] &= \log\Gamma(\sum_{j=1}^{K}\alpha_{D,j}\beta^*) - \sum_{i=1}^{K}\log\Gamma(\alpha_{D,i}\beta^*) \\
&\quad + \sum_{i=1}^{K}(\alpha_{D,i}\beta^* - 1)\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)\right). \quad (5)
\end{aligned}$$

Computing the entropy for the variational distribution terms which are specific to a single document is relatively straightforward:

$$\begin{aligned}
\mathbb{E}_q\left[\log q(z)\right] &= \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\log\phi_{n,i} \\
\mathbb{E}_q\left[\log q(\theta)\right] &= \log\Gamma(\sum_{j=1}^{K}\gamma_j) + \sum_{i=1}^{K}\log\Gamma(\gamma_i) - \sum_{i=1}^{K}(\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)\right)
\end{aligned}$$

We now have all the terms necessary to express the likelihood bound contribution of a document

$$\begin{aligned}
L_d &= \log\Gamma(\sum_{j=1}^{K}\alpha_{D,j}\beta^*) - \sum_{i=1}^{K}\log\Gamma(\alpha_{D,i}\beta^*) \\
&\quad + \sum_{i=1}^{K}(\alpha_{D,i}\beta^* - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)) \\
&+ \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)\right) + \sum_{n=1}^{N}\sum_{i=1}^{K}\sum_{j=1}^{K}\phi_{n,i}\phi_{p(n),j}\left(\Psi(\nu_{j,i}) - \Psi(\sum_{k=1}^{K}\nu_{j,k})\right) \\
&\quad - \left(\sum_{n=1}^{N}\omega_n^{-1}\sum_{i=1}^{K}\sum_{j=1}^{K}\phi_{p(n),j}\frac{\gamma_i\nu_{j,i}}{\sum_{k=1}^{K}\gamma_k\sum_{k=1}^{K}\nu_{j,k}} + \log\omega_n - 1\right) \\
&+ \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_i\log\tau_{i,w_n} \\
&- \log\Gamma(\sum_{j=1}^{K}\gamma_j) + \sum_{i=1}^{K}\log\Gamma(\gamma_i) - \sum_{i=1}^{K}(\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)\right) \\
&- \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\log\phi_{n,i}. \qquad (6)
\end{aligned}$$

3.2 Variational updates

Because we seek to maximize Equation 6, we compute the gradient of the likelihood bound with respect to each variable. When possible, we'll explicitly solve for a value of the parameter that maximizes the likelihood with respect to that coordinate. Otherwise, we'll fully express the likelihood function and gradient for that variational parameter so that we can use these expressions in a numerical optimization package. The variational parameters are presented in an order corresponding roughly to the order in which they are used by the algorithm.

3.2.1 Topic normalizer slack variable

First, we'll explicitly solve for the value of ω, the slack variable introduced in the Taylor approximation of Equation 3.

$$\begin{aligned}
\frac{\partial L}{\partial\omega_n} &= \omega_n^{-2}\left(\sum_{i=1}^{K}\sum_{j=1}^{K}\phi_{p(n),j}\frac{\gamma_i\nu_{j,i}}{\sum_{k=1}^{K}\gamma_k\sum_{k=1}^{K}\nu_{j,k}}\right) - \omega_n^{-1} \qquad (7) \\
\Rightarrow \omega_n &= \sum_{i=1}^{K}\sum_{j=1}^{K}\phi_{p(n),j}\frac{\gamma_i\nu_{j,i}}{\sum_{k=1}^{K}\gamma_k\sum_{k=1}^{K}\nu_{j,k}}
\end{aligned}$$

3.2.2 Variational topic multinomial

Because ϕ is a variational multinomial, it must sum to one. So we want to maximize the likelihood bound augmented by that constraint and its corresponding Lagrange multiplier.

$$
\begin{aligned}
L_{\phi_{ni}} &= L_d + \lambda \left(\sum_{j=1}^{K} \phi_{n,j} - 1 \right) && (8) \\
\frac{\partial L_{\phi_{ni}}}{\partial \phi_{ni}} &= \Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j) + \sum_{j=1}^{K} \phi_{p(n),j} \left(\Psi(\nu_{j,i}) - \Psi(\sum_{k=1}^{K} \nu_{j,k}) \right) \\
&+ \sum_{c \in c(n)} \sum_{j=1}^{K} \phi_{c,j} \left(\Psi(\nu_{i,j}) - \Psi(\sum_{k=1}^{K} \nu_{i,k}) \right) \\
&- \sum_{c \in c(n)} \omega_c^{-1} \sum_{j}^{K} \frac{\gamma_j \nu_{i,j}}{\sum_k \gamma_k \sum_k \nu_{i,k}} \\
&+ \phi_i \log \tau_{i,w_n} \\
&- \log \phi_{n,i} \\
&+ \lambda. && (9)
\end{aligned}
$$

This then gives us, up to a constant, the value of ϕ that maximizes our variational bound

$$
\begin{aligned}
\phi_{ni} \propto \exp \Bigg\{ & \Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j) + \sum_{j=1}^{K} \phi_{p(n),j} \left(\Psi(\nu_{j,i}) - \Psi(\sum_{k=1}^{K} \nu_{j,k}) \right) \\
& - \sum_{c \in c(n)} \omega_c^{-1} \sum_{j}^{K} \frac{\gamma_j \nu_{i,j}}{\sum_k \gamma_k \sum_k \nu_{i,k}} \\
& + \sum_{c \in c(n)} \sum_{j=1}^{K} \phi_{c,j} \left(\Psi(\nu_{i,j}) - \Psi(\sum_{k=1}^{K} \nu_{i,k}) \right) \\
& + \log \tau_{i,w_n} \Bigg\}. \qquad (10)
\end{aligned}
$$

3.2.3 Variational document Dirichlet

Because we couple π and θ, the interaction between these terms in the normalizer prevents us from solving the optimization explicitly. But we still need to derive the functional form and the derivative for the conjugate gradient algorithm [1]. In doing so, the following derivative is useful:

$$
\begin{aligned}
f &= \sum_{i=1}^{N} \alpha_i \frac{x_i}{\sum_{i=j}^{N} x_j} \\
\Rightarrow \frac{\partial f}{\partial x_i} &= \frac{\alpha_i \sum_{j \neq i}^{N} x_j - \sum_{j \neq i}^{N} \alpha_j x_j}{\left(\sum_{i=1}^{N} x_i \right)^2}. \qquad (11)
\end{aligned}
$$

First, we write the terms in L_d that involve γ:

$$\begin{aligned} L_\gamma &= \sum_{i=1}^{K} (\alpha_{D,i}\beta^* - 1)\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)\right) \\ &+ \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)\right) \\ &- \left(\sum_{n=1}^{N}\omega_n^{-1}\sum_{i=1}^{K}\sum_{j=1}^{K}\phi_{p(n),j}\frac{\gamma_i\nu_{j,i}}{\sum_{k=1}^{K}\gamma_k\sum_{k=1}^{K}\nu_{j,k}} + \log\omega_n - 1\right) \\ &- \log\Gamma(\sum_{j=1}^{K}\gamma_j) + \sum_{i=1}^{K}\log\Gamma(\gamma_i) - \sum_{i=1}^{K}\left((\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)\right)\right). \end{aligned} \tag{12}$$

Now, we use Equation 11 to compute the partial derivative with respect to γ_i to be

$$\begin{aligned} \frac{\partial L}{\partial\gamma_i} &= \Psi'(\gamma_i)\left(\alpha_{D,i}\beta^* + \sum_{n=1}^{N}\phi_{n,i} - \gamma_i\right) - \Psi'(\sum_{j=1}^{N}\gamma_j)\sum_{j=1}^{K}\left[\alpha_{D,j}\beta^* + \sum_{n=1}^{N}\phi_{n,j} - \gamma_j\right] \\ &- \sum_{n=1}^{N}\omega_n^{-1}\sum_{j=1}^{K}\left[\phi_{p(n),j}\frac{\nu_{j,i}\sum_{k\neq j}^{N}\gamma_k - \sum_{k\neq j}^{N}\nu_{j,k}\gamma_k}{\left(\sum_{k=1}^{N}\gamma_k\right)^2\sum_{k=1}^{N}\nu_{j,k}}\right]. \end{aligned} \tag{13}$$

4 Global variational terms

In addition to the terms that are specific to the words of a single document, we also have parameters that deal with the word probabilities within topics, the transitions between topics, and the top-level topic weights. We return to these terms and, after deriving the rest of the likelihood bound, describe the updates for each of these global parameters.

4.1 Expanding expectations

Let's first consider the terms associated with the variational Dirichlet associated with the topic transitions from parent to child in the parse tree, which gives us

$$\begin{aligned} \mathbb{E}_q\left[p(\pi_d|\alpha_P,\beta)\right] &= \log\Gamma(\sum_{j=1}^{K}\alpha_{T,j}\beta^*) - \sum_{i=1}^{K}\log\Gamma(\alpha_{T,i}\beta^*) \\ &+ \sum_{i=1}^{K}(\alpha_{T,i}\beta^* - 1)(\Psi(\nu_i) - \Psi(\sum_{j=1}^{K}\nu_j)) \end{aligned} \tag{14}$$

$$\begin{aligned} \mathbb{E}_q\left[\log q(\pi)\right] &= \log\Gamma(\sum_{j=1}^{K}\nu_j) + \sum_{i=1}^{K}\log\Gamma(\nu_i) \\ &- \sum_{i=1}^{K}(\nu_i - 1)\left(\Psi(\nu_i) - \Psi(\sum_{j=1}^{K}\nu_j)\right) \end{aligned} \tag{15}$$

Another global term that we have not considered is for the top-level weights associated with β. The distribution is defined in terms of stick breaking proportions [2], so, following the derivation of Liang et al [3], we must transform $\beta*$ into these stick breaking proportions; this is accomplished by dividing each weight by the tail sum

$$T_z \equiv 1 - \sum_{i}^{z-1}\beta_i, \tag{16}$$

the sum of the remaining weights. This also introduces an extra term for the change of variables.

$$\begin{aligned}
\mathbb{E}_q\left[p(\beta|\alpha)\right] &= \mathbb{E}_q\left[\log \text{GEM}(\beta;\alpha)\right] \\
&= \log\left(\prod_{z=1}^{K-1} \text{Beta}(\beta^*/T_z;1,\alpha)\right) - \sum_{z}^{K-1}\log T_z \\
&= \sum_{z=1}^{K-1}(\alpha-1)\log\left(\frac{T_{z+1}}{T_z}\right) + (K-1)\log\alpha - \sum_{z}^{K-1}\log T_z \\
&= (\alpha-1)\log T_K - \sum_{z}^{K-1}\log T_z + C. \qquad (17)
\end{aligned}$$

Lastly, the multinomial distribution over vocabulary terms per topic is straightforward

$$\mathbb{E}_q\left[\log p(\boldsymbol{\tau}|\sigma)\right] = \sum_{i=1}^{K}\sum_{v=1}^{V}\sigma\log\tau_{i,v}. \qquad (18)$$

The total likelihood is, after including Equation 6, then

$$\begin{aligned}
L &= \sum_{d}^{M} L_d \\
&+ (\alpha-1)\log T_K - \sum_{z}^{K-1}\log T_z \\
&+ \log\Gamma(\sum_{j=1}^{K}\alpha_{T,j}\beta^*) - \sum_{i=1}^{K}\log\Gamma(\alpha_{T,i}\beta^*) + \sum_{i=1}^{K}(\alpha_{T,i}\beta^*-1)(\Psi(\nu_i) - \Psi(\sum_{j=1}^{K}\nu_j)) \\
&- \log\Gamma(\sum_{j=1}^{K}\nu_j) + \sum_{i=1}^{K}\log\Gamma(\nu_i) - \sum_{i=1}^{K}(\nu_i-1)\left(\Psi(\nu_i) - \Psi(\sum_{j=1}^{K}\nu_j)\right) \\
&+ \sum_{i=1}^{K}\sum_{v=1}^{V}\sigma\log\tau_{i,v}. \qquad (19)
\end{aligned}$$

4.2 Variational updates

These are the updates which require input from all documents and cannot be parallelized. However, each of the document-specific updates can compute sufficient statistics for these computations which are summed together for the final updates.

4.2.1 Variational transition Dirichlet

Like our update for γ, the interaction between π and θ in the normalizer prevents us from solving the optimization explicitly. First, consider the variational Dirichlet for transitions from topic i

$$\begin{aligned}
L_{\nu_i} &= \sum_{j=1}^{K}(\alpha_{P,j}-1)(\Psi(\nu_{i,j}) - \Psi(\sum_{k=1}^{K}\nu_{i,k})) \\
&+ \sum_{n=1}^{N}\sum_{c\in c(n)}\sum_{j=1}^{K}\phi_{n,i}\phi_{c,j}\left(\Psi(\nu_{i,j}) - \Psi(\sum_{k=1}^{K}\nu_{i,k})\right) \\
&- \left(\sum_{n=1}^{N}\phi_{n,i}\sum_{c\in c(n)}\left[\omega_c^{-1}\sum_{j=1}\frac{\gamma_j\nu_{i,j}}{\sum_{k=1}^{K}\gamma_k\sum_{k=1}^{K}\nu_{i,k}}\right] + \log\omega_n - 1\right) \\
&- \log\Gamma(\sum_{j=1}^{K}\nu_{i,j}) + \sum_{j=1}^{K}\log\Gamma(\nu_{i,j}) - \sum_{j=1}^{K}(\nu_{i,j}-1)\left(\Psi(\nu_{i,j}) - \Psi(\sum_{k=1}^{K}\nu_{i,k})\right). \qquad (20)
\end{aligned}$$

Differentiating this expression, keeping in mind Equation 11, gives

$$\begin{aligned}\frac{\partial L}{\partial \nu_{i,j}} &= \Psi'(\nu_{i,j})\left(\alpha_{P,j} + \sum_{n=1}^{N}\sum_{c\in c(n)}\phi_{n,i}\phi_{c,j} - \nu_{i,j}\right) \\ &\quad -\Psi'(\sum_{k=1}^{K}\nu_{i,k})\sum_{k=1}^{K}\left[\alpha_{P,k} + \sum_{n=1}^{N}\sum_{c\in c(n)}\phi_{n,i}\phi_{c,k} - \nu_{i,k}\right] \\ &\quad -\sum_{n}^{N}\phi_{n,i}\sum_{c\in c(n)}\left[\omega_c^{-1}\frac{\gamma_j\sum_{k\neq j}^{N}\nu_{i,k} - \sum_{k\neq j}^{N}\nu_{i,k}\gamma_k}{\left(\sum_{k=1}^{N}\nu_{j,k}\right)^2\sum_{k=1}^{N}\gamma_k}\right], \end{aligned} \tag{21}$$

which is sufficient for conjugate gradient optimization [1].

4.2.2 DP Process

The last variational parameter is $\beta*$, which is the variational estimate of the top-level weights β. Note that we separate the final term β_K and the previous weights; this is because β_K is implicitly defined as $\left(1 - \sum_{i=0}^{K-1}\beta_i\right)$.

$$\begin{aligned}L_{\beta^*} &= (\alpha-1)\log T_K - \sum_{z}^{K-1}\log T_z \\ &+ \sum_{d=1}^{M}\left[\log\Gamma(\alpha_D) - \sum_{i}^{K-1}\log\Gamma(\alpha_D\beta_i^*) - \log\Gamma(\alpha_D\beta_K^*)\right. \\ &\quad \left.+\sum_{i=1}^{K}(\alpha_D\beta_i^* - 1)\left(\Psi(\gamma_{d,i}) - \Psi(\sum_{j=1}^{K}\gamma_{d,j})\right)\right] \\ &+ \sum_{k=1}^{K}\left[\log\Gamma(\alpha_T) - \sum_{i}^{K-1}\log\Gamma(\alpha_T\beta_i^*) - \log\Gamma(\alpha_T\beta_K^*)\right. \\ &\quad \left.+\sum_{i=1}^{K}(\alpha_T\beta_i^* - 1)\left(\Psi(\nu_{k,i}) - \Psi(\sum_{j=1}^{K}\nu_{k,j})\right)\right]. \end{aligned} \tag{22}$$

Taking the derivative of this expression, and implicitly differentiating β_K gives us

$$\frac{\partial L_{\beta^*}}{\partial \beta_k^*} = \left(\sum_{z=k+1}^{K-1}\frac{1}{T_z}\right) - \frac{\alpha-1}{T_K} \tag{23}$$

$$\begin{aligned}&+ \alpha_D\sum_{d}^{M}\left(\Psi(\gamma_{d,k}) - \Psi(\sum_{j=1}^{K}\gamma_{d,j})\right) - \alpha_D\sum_{d}^{M}\left(\Psi(\gamma_{d,K}) - \Psi(\sum_{j=1}^{K}\gamma_{d,j})\right) \\ &+ \alpha_T\sum_{z}^{K}\left(\Psi(\nu_{z,k}) - \Psi(\sum_{j=1}^{K}\nu_{z,j})\right) - \alpha_T\sum_{z}^{K}\left(\Psi(\nu_{z,K}) - \Psi(\sum_{j=1}^{K}\nu_{z,j})\right) \\ &- K\left[\alpha_T\Psi(\alpha_T\beta_k^*) - \alpha_T\Psi(\alpha_T\beta_K^*)\right] \\ &- M\left[\alpha_D\Psi(\alpha_D\beta_k^*) - \alpha_D\Psi(\alpha_D\beta_K^*)\right] \end{aligned} \tag{24}$$

which we optimize through the barrier constrained optomization algorithm [4].

References

[1] Galassi, M., J. Davies, J. Theiler, et al. *Gnu Scientific Library: Reference Manual*. Network Theory Ltd., 2003.

[2] Pitman, J. Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11:501–514, 2002.

[3] Liang, P., S. Petrov, M. Jordan, et al. The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of Emperical Methods in Natural Language Processing*, pages 688–697. 2007.

[4] Boyd, S., L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.