1 Appendix

1.1 Variational posterior and evidence lower bound

In this section, we describe the evidence lower bound and expand its terms to derive the variational updates in the following section. The evidence is given by the following formula:

$$
\begin{aligned}
\mathcal{L}(q) &= \log p(d_{1:T}) && (1)\\
&\geq \int q(\beta, l, \theta, z | \tilde{\beta}, \tilde{l}, \gamma, \phi) \log \left(\frac{p(\beta, l, \theta, z) p(d | \beta, l, \theta, z)}{q(\beta, l, \theta, z | \tilde{\beta}, \tilde{l}, \gamma, \phi)} \right) d_{\beta_{1:T}} && (2)\\
&= \mathbb{E}_q \left[\log \prod_T \prod_K p(l_{T,k}) \right] && (3)\\
&+ \mathbb{E}_q \left[\log \prod_T \prod_{D_t} \prod_{N_{d_t}} p(z_n | \theta_{d_t}) \right] && (4)\\
&+ \mathbb{E}_q \left[\log \prod_{t=1}^{T} \prod_K p(\beta_{t,k} | \beta_{t-f,k}) \right] && (5)\\
&+ \mathbb{E}_q \left[\log \prod_T \prod_{D_t} \prod_{N_{d_t}} p(w_n | z_n) \right] && (6)\\
&+ H(q) && (7)\\
&+ \ldots, && (8)
\end{aligned}
$$

where we have left out some terms (8) which are not relevant to this model's derivation. To maximize this lower bound, we find locally optimal values for the parameters $\phi, \tilde{\beta}, \tilde{l}$, and γ numerically through the variational updates described in the following section. We expand these terms and derive the updates in the supplementary material.

We can expand 3 as:

$$
\begin{aligned}
\mathbb{E}_q \left[\log \prod_T \prod_K p(l_{T,k}) \right] &= \sum_T \sum_{D_t} \sum_K \mathbb{E}_q \left[-\frac{l_{d,k}^2}{2\sigma_d^2} - \frac{1}{2}(\log 2\pi + \log \sigma_d^2) \right] && (9)\\
&= \sum_T \sum_{D_t} \sum_K -\frac{1}{2\sigma_d^2} (\tilde{l}_{d_t,k}^2 + \sigma_l^2) - \frac{1}{2}(\log 2\pi + \log \sigma_d^2) && (10)\\
& && (11)
\end{aligned}
$$

Equation 4 can be expanded as demonstrated in [**?**]:

$$
\begin{aligned}
\mathbb{E}_q \left[\log \prod_T \prod_{D_t} \prod_{N_{d_t}} p(z_n | \theta_{d_t}) \right] &= \sum_T \sum_{D_t} \sum_{N_{d_t}} \mathbb{E}_q \left[\log p(\mathbf{z}_{d_t} | \theta_{d_t}) \right] && (12)\\
&= \sum_N \sum_K \phi_{n,k} \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j) \right) && (13)\\
& && (14)
\end{aligned}
$$

Title	Year	Doc Citations	Median
Sentence Generation By Semantic Concordance	1965-01-01	0	0
Man-Aided Computer, Translation From English Into French Using An On-Line System To Manipulate A Bi-Lingual Conceptual Dictionary, Or Thesaurus	1967-01-01	0	0
On The Use Of Linguistic Quantifying Operators In The Logico-Semantic Structure Representation Of Utterances	1969-01-01	0	0
Computational Linguistics And Linguistic Theory	1973-01-01	0	0
Stereotypes As An Actor Approach Towards Solving The Problem Of Procedural Attachment In Frame Theories	1975-01-01	0	1
The Processing Of Referring Expressions Within A Semantic Network	1978-01-01	0	1
Simple Digital Speech Synthesis	1979-01-01	0	1
Interactive Discourse: Looking To The Future Panel Chair's Introduction	1980-01-01	0	0
Discourse-Oriented Anaphora Resolution In Natural Language Understanding: A Review	1981-01-01	**6**	1
Natural-Language Interface	1982-01-01	**2**	0
How To Parse Gaps In Spoken Utterances	1983-01-01	1	1
A Stochastic Approach To Sentence Parsing	1984-01-01	**4**	2
Using Restriction To Extend Parsing Algorithms For Complex-Feature-Based Formalisms	1985-01-01	**57**	3
On The Use Of Term Associations In Automatic Information Retrieval	1986-01-01	1	1
Tools And Methods For Computational Lexicology	1987-01-01	**22**	1
The Experience Of Developing A Large-Scale Natural Language Text Processing System: Critique	1988-01-01	**4**	2
Improvements In The Stochastic Segment Model For Phoneme Recognition	1989-01-01	**3**	2
Deducing Linguistic Structure From The Statistics Of Large Corpora	1990-01-01	**22**	1
A Dynamic Language Model For Speech Recognition	1991-01-01	**11**	1
Feature Selection And Feature Extract Ion For Text Categorization	1992-01-01	**3**	1
HMM-Based Part-Of-Speech Tagging For Chinese Corpora	1993-01-01	**7**	1
Similarity-Based Estimation Of Word Cooccurrence Probabilities	1994-01-01	**24**	1
Text Chunking Using Transformation-Based Learning	1995-01-01	**143**	4
A Maximum Entropy Model For Part-Of-Speech Tagging	1996-01-01	**215**	1
High Performance Segmentation Of Spontaneous Speech Using Part Of Speech And Trigger Word Information	1997-01-01	2	2
Trainable, Scalable Summarization Using Robust NLP And Machine Learning	1998-01-01	**5**	1
Untangling Text Data Mining	1999-01-01	**8**	2
Use Of Support Vector Learning For Chunk Identification	2000-01-01	**38**	2
Low-Cost, High-Performance Translation Retrieval: Dumber Is Better	2001-01-01	2	2
Extracting Important Sentences With Support Vector Machines	2002-01-01	**5**	2
Evaluation And Extension Of Maximum Entropy Models With Inequality Constraints	2003-01-01	**10**	2
Text Mining - Next Steps For Drug Discovery	2004-01-01	0	1
A General Technique To Train Language Models On Language Models	2005-01-01	2	2

Figure 1: The most influential documents found by the DIM in each year of the ACL corpus. The number of citations of each document is given in the Citations column, and median and third quartile citations of all documents for the same date are given for reference. Omitting the all-zero columns and splitting ties, these documents are above the median 76% of the time and above the third quartile 41% of the time.

Finally, we expand 5:

$$
\begin{aligned}
\mathbb{E}_q\left[\log\prod_{t=1}^{T}\prod_{K} p(\beta_{t,k}|\beta_{t-f,k})\right] =& \sum_{t=1}^{T}\sum_{K}\sum_{W} -\frac{1}{2\sigma^2}\mathbb{E}_q\left[\beta_{t,k,w}^2+\beta_{t-f,k,w}^2\right] \qquad (15)\\
&+\frac{1}{\sigma^2}\mathbb{E}_q\left[\beta_{t,k,w}\beta_{t-f,k,w}\right]\\
&-\frac{1}{\sigma^2}\mathbb{E}_q\left[(\beta_{t-f,k,w}-\beta_{t,k,w})\circ\exp(-\beta_{t-f,k,w})(\mathbf{W}_{t-f,w}\circ[z_w]_k)l_{t-f,k}\right]\\
&+\frac{1}{\sigma^2}\mathbb{E}_q\left[\exp(-2\beta_{t-f,k,w})\left((\mathbf{W}_{t-f,k,w}\circ[z_w]_k)l_{t-2,k}\right)^2\right]\\
&-\frac{VT}{2}(\log\sigma^2+\log 2\pi)\\
=&-\frac{VT}{2}(\log\sigma^2+\log 2\pi)\\
&-\sum_{t=1}^{T}\frac{1}{2\sigma^2}(\tilde{m}_t-\tilde{m}_{t-f})^2\\
&-\frac{1}{\sigma^2}\sum_{t=1}^{T}\mathrm{Tr}(\tilde{V}_t)+\frac{1}{2\sigma^2}\left(\mathrm{Tr}(\tilde{V}_0)-\mathrm{Tr}(\tilde{V}_T)\right)\\
&+\frac{1}{\sigma^2}\exp(-\tilde{m}_{t-f,k}+\tilde{V}_{t,k}/2)^T(\tilde{m}_t-\tilde{m}_{t-f}+\tilde{V}_{t-f})\circ(\mathbf{W}_{t-f,k}\circ\phi_{t-f,k})\tilde{l}_{t-f,k} \qquad (16)\\
&-\frac{1}{2\sigma^2}\left((\mathbf{W}_{t-f,k}\circ\phi_{t-f,k})l_{t-f,k}\right)^T\Lambda_{\exp(-2\tilde{m}_{t-f,k}+2\tilde{V}_{t,k})}\left((\mathbf{W}_{t-f,k}\circ\phi_{t-f,k})l_{t-f,k}\right)^T\\
&-\frac{1}{2\sigma^2}\exp(-2\tilde{m}_{t-f,k}+2\tilde{V}_{t,k})^T(\mathbf{W}_{t-f,k}\circ\mathbf{W}_{t-f,l}\circ(\phi_{t-f,k}-\phi_{t-f,k}\circ\phi_{t-f,k}))(\tilde{l}_{t-f,k}\circ\\
&-\frac{1}{2\sigma^2}\exp(-2\tilde{m}_{t-f,k}+2\tilde{V}_{t,k})^T(\mathbf{W}_{t-f,k}\circ\mathbf{W}_{t-f,k}\circ\phi_{t-f,k}\circ\phi_{t-f,k})\vec{\sigma_l^2}_{D_{t-f}}.
\end{aligned}
$$

Above, $\circ$ refers to the Hadamard element-wise product and $\Lambda_{\vec{x}}$ refers to a diagonal matrix having the elements of $\vec{x}$ on its diagonal.

Further, on line 16, we have used the fact that $\mathbb{E}_q\left[\beta_t\exp(-\beta_t)\right]=(\tilde{m}-\tilde{V})\exp(-\tilde{m}+\tilde{V}/2)$.

1.2 Update equations

We update θ and as in the DTM. The updates for $\tilde{\beta}$ and ϕ are different in the Document Influence Model, and the document weights $\tilde{l}$ must also be updated. As shown in equation **??**, the document weights are updated with a regression. We determine this regression by collecting terms with $\tilde{l}$, taking the derivative, and setting equal to zero.

To find the updates for ϕ, we gather all terms from the evidence lower bound containing ϕ and form the Lagrangian to enforce the constraint $\sum_{j=1}^{K} \phi_{n,j} = 1$:

$$
\begin{aligned}
L[\phi] = \sum_N \sum_K \Bigg(\phi_{n,k} \bigg(& - \log \phi_{n,k} + (\Psi(\gamma_k) - \Psi(\sum_{j=1}^{K} \gamma_j)) + \tilde{m}_{n,k} \bigg) \\
& + \lambda_n (\sum_{j=1}^{K} \phi_{n,j} - 1) \\
& + \frac{1}{\sigma^2} (\tilde{m}_{t+f} - \tilde{m}_t + \tilde{V}_t) \exp(-\tilde{m} + \tilde{V}/2) \phi_{n,k} \tilde{l}_{t,k,d_n} w_{t,d,n} \\
& - \frac{1}{\sigma^2} \exp(-2\tilde{m}_{t,n,k} + 2\tilde{V}_{t,n,k}) (w_{t,d_n,n} \tilde{l}_{t,k,d_n} \phi_{n,k} \sum_{d' \neq d_n} (w_{t,d',n} \tilde{l}_{t,k,d'} \phi_{n'_d,k})) \\
& - \frac{1}{2\sigma^2} \exp(-2\tilde{m}_{t,n,k} + 2\tilde{V}_{t,n,k}) (w^2_{t,d_n,n} \phi_{n,k} (\tilde{l}^2_{t,k,d_n} + \sigma_l^2)) \Bigg)
\end{aligned}
$$

Next, take the derivative with respect to $\phi_{n,i}$:

$$
\begin{aligned}
\frac{\partial L}{\partial \phi_{n,k}} = & - \log \phi_{n,k} - 1 + \Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j) + \lambda_n \qquad (17) \\
& + \frac{1}{\sigma^2} (\tilde{m}_{t+f} - \tilde{m}_t + \tilde{V}_t) \exp(-\tilde{m} + \tilde{V}/2) \tilde{l}_{t,k,d_n} w_{t,d,n} \\
& - \frac{1}{\sigma^2} \exp(-2\tilde{m}_{t,n,k} + 2\tilde{V}_{t,n,k}) w_{t,d_n,n} \tilde{l}_{t,k,d_n} \left(\sum_D (w_{t,d,n} \tilde{l}_{t,k,d}) - 2\phi_{n,k} w_{t,d,n} \tilde{l}_{t,k,d} \right) \\
& - \frac{1}{2\sigma^2} \exp(-2\tilde{m}_{t,n,k} + 2\tilde{V}_{t,n,k}) (w^2_{t,d_n,n} (\tilde{l}^2_{t,k,d_n} + \sigma_l^2)) \Bigg)
\end{aligned}
\qquad (18)
$$

Observing the ϕ and $\log \phi$ terms on the RHS of 17 means that we cannot necessarily update ϕ exactly in a single pass, as in LDA. Once the variables ϕ are updated as in Equation **??**, we can then select $\lambda_{n,s}$ to minimize the sum of squares $\sum_K \left(\frac{\partial L}{\partial \phi_{n,k}} \right)^2$ of these partial derivatives – which are not necessarily zero anymore because ϕ has been renormalized.

The update for $\tilde{\beta}$ can be found by collecting terms containing $\tilde{\beta}$ from Equation 1. We then maximize with respect to $\tilde{\beta}$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \tilde{\beta}_{sw}} =& -\frac{1}{\sigma^2}\sum_{t=1}^{T}(\tilde{m}_{tw}-\tilde{m}_{t-f,w})\left(\frac{\partial \tilde{m}_{tw}}{\partial \tilde{\beta}_{sw}}-\frac{\partial \tilde{m}_{t-f,w}}{\partial \tilde{\beta}_{sw}}\right)\\
&+\sum_{T}\left(n_{tw}-n_t\zeta^{-1}\exp(\hat{m}_{\beta_{tw}}+\frac{\tilde{V}_{tw}}{2})\right)\frac{\partial \tilde{m}_t}{\partial \tilde{\beta}_{sw}}\\
&+\frac{\exp(-\tilde{m}_{t-f,w}+\tilde{V}_{t-f,w}/2)}{\sigma^2}\left((-\tilde{m}_{t,w}+\tilde{m}_{t-f,w}-\tilde{V}_{t-f,w}-1)\frac{\partial \tilde{m}_{t-f,w}}{\partial \tilde{\beta}_{sw}}+\frac{\partial \tilde{m}_{tw}}{\partial \tilde{\beta}_{sw}}\right)\\
&\times(\mathbf{W}_{t-f,k}\circ\phi_{t-f,k})\tilde{l}_{t-f,k}\\
&+\frac{\exp(-2\tilde{m}_{t-f,w}+2\tilde{V}_{t-f,w})}{\sigma^2}\frac{\partial \tilde{m}_{t-f,w}}{\partial \tilde{\beta}_{sw}}\\
&\times(((\mathbf{W}_{t-f,k}\circ\phi_{t-f,k})\tilde{l}_{t-f,k})^2\\
&+(\mathbf{W}_{t-f,k}\circ\mathbf{W}_{t-f,k}\circ(\phi_{t-f,k}-\phi_{t-f,k}\circ\phi_{t-f,k}))\tilde{l}_{t-f,k}\circ\tilde{l}_{t-f,k}+\vec{\sigma_l^2})\\
&+(\mathbf{W}_{t-f,k}\circ\mathbf{W}_{t-f,k}\circ\phi_{t-f,k}\circ\phi_{t-f,k})\vec{\sigma_l^2})
\end{aligned}$$