9 Appendix

9.1 Code Base

`https://github.com/jasonren12/PredictionFocusedTopicModel`

9.2 Theorem Proofs

Theorem 1. *Suppose that the channel switches ξ_d and the document topic distribution θ_d are conditionally independent in the posterior for all documents, then β and π have disjoint supports over the vocabulary.*

Proof. For simplicity of notation, we assume a single document and hence drop the subscripts on ξ_d and θ_d. All of the arguments are the same in the multi-document case. If ξ and θ are conditionally independent in the posterior, then we can factor the posterior as follows: $p(\xi, \theta|\mathbf{w}, y) = p(\xi|\mathbf{w}, y)p(\theta|\mathbf{w}, y)$. We expand out the posterior:

$$\begin{aligned} p(\xi, \theta|\mathbf{w}, y) &\propto p(\xi)p(\theta)p(\mathbf{w}, y|\theta, \xi) \\ &\propto p(\xi)p(\theta)p(y|\theta)\prod_n p_\beta(w_n|\theta)^{\xi_n} p_\pi(w_n)^{1-\xi_n} \\ &= f(\theta)g(\xi)\prod_n p_\beta(w_n|\theta)^{\xi_n} \end{aligned}$$

for some functions f and g. Thus we see that we must have that $\prod_n p_\beta(w_n|\theta)^{\xi_n}$ factors into some $r(\theta)s(\xi)$. We expand $\prod_n p_\beta(w_n|\theta)^{\xi_n}$:

$$\begin{aligned} p_\beta(w_n|\theta)^{\xi_n} &= \left(\sum_k \beta_{k,w_n}\theta_k\right)^{\xi_n} \\ &= I(\xi_n = 0) + I(\xi_n = 1)\left(\sum_k \beta_{k,w_n}\theta_k\right) \end{aligned}$$

So that we can express the product as:

$$\prod_n p_\beta(w_n|\theta)^{\xi_n} = \prod_n \left\{ I(\xi_n = 0) + I(\xi_n = 1)\left(\sum_k \beta_{k,w_n}\theta_k\right)\right\}$$

In order to further simplify, let $\beta_0 = \{n : \sum_k \beta_{k,w_n} = 0\}$ and $\beta_> = \{n : \sum_k \beta_{k,w_n} > 0\}$. In other words β_0 is the set of n such that the word w_n is not supported by β, and $\beta_>$ is the set of n such that the word w_n is supported by β.

We can rewrite the above as:

$$\prod_n p_\beta(w_n|\theta)^{\xi_n} = \left(\prod_{n\in\beta_0} I(\xi_n = 0)\right)\left(\prod_{n\in\beta_>}\left\{I(\xi_n = 0) + I(\xi_n = 1)\sum_k \beta_{k,w_n}\theta_k\right\}\right)$$

Thus, we see that we can factor $\prod_n p_\beta(w_n|\theta)^{\xi_n}$ as a function of θ and ξ into the form $r(\theta)s(\xi)$ only if $\xi_n = 0$ or $\xi_n = 1$ with probability 1. We can check that this implies $\beta_k^\top \pi = 0$ for each k by the result of Theorem 2. □

Theorem 2. $\beta^\top \pi = 0$ *if and only if there exists a* $\boldsymbol{\xi}^*$ *s.t.* $p(\boldsymbol{\xi}^*|\mathbf{w}, y) = 1$

Proof.
1. Assume $\beta^\top \pi = 0$. Then, conditional on w_n, $\xi_n = 1$ with probability 1 if $\pi_{w_n} = 0$ and $\xi_n = 0$ with probability 1 if $\pi_{w_n} > 0$. So we have $p(\boldsymbol{\xi}^*|\mathbf{w}, y) = 1$ for the $\boldsymbol{\xi}^*$ corresponding to $\mathbf{w}$ as described before.

2. Assume there exists a $\boldsymbol{\xi}^*$ s.t. $p(\boldsymbol{\xi}^*|\mathbf{w}, y) = 1$.

   Then we have:

$$\begin{aligned} p(\boldsymbol{\xi}^*|\mathbf{w}, y) &= \frac{p(\mathbf{w}, y|\boldsymbol{\xi}^*)p(\boldsymbol{\xi}^*)}{\sum_{\boldsymbol{\xi}} p(\mathbf{w}, y|\boldsymbol{\xi})p(\boldsymbol{\xi})} = 1 \\ p(\mathbf{w}, y|\boldsymbol{\xi}^*)p(\boldsymbol{\xi}^*) &= \sum_{\boldsymbol{\xi}} p(\mathbf{w}, y|\boldsymbol{\xi})p(\boldsymbol{\xi}) \end{aligned}$$

   This implies $p(\mathbf{w}, y|\boldsymbol{\xi})p(\boldsymbol{\xi}) = 0 \; \forall \; \boldsymbol{\xi} \neq \boldsymbol{\xi}^*$, which implies $p(\mathbf{w}, y|\boldsymbol{\xi}) = 0 \; \forall \; \boldsymbol{\xi} \neq \boldsymbol{\xi}^*$

   Then we have:

$$\begin{aligned} p(\mathbf{w}, y|\boldsymbol{\xi}) &= p(y|\mathbf{w}, \boldsymbol{\xi})p(\mathbf{w}|\boldsymbol{\xi}) \\ &= \left(\int_\theta p(y|\theta)p(\theta|\mathbf{w}, \boldsymbol{\xi})d\theta\right)\left(\int_\theta p(\mathbf{w}|\theta, \boldsymbol{\xi})p(\theta)d\theta\right) \end{aligned}$$

   The first term will be greater than 0 because $y|\theta$ is distributed Normal. We focus on the second term.

$$\int_\theta p(\mathbf{w}|\theta, \boldsymbol{\xi})p(\theta)d\theta = \int_\theta p(\theta)\prod_n p_\beta(w_n|\theta)^{\xi_n} p_\pi(w_n)^{1-\xi_n} d\theta$$

   Let X be the set of $\boldsymbol{\xi}$ that differ from $\boldsymbol{\xi}^*$ in one and only one position, i.e. $\xi_n = \xi_n^*$ for all $n \in \{1, \ldots N\} \setminus \{i\}$ and $\xi_i \neq \xi_i^*$. For each $\xi \in X$, $\int_\theta d\theta p(\theta)\prod_n p_\beta(w_n|\theta)^{\xi_n} p_\pi(w_n)^{1-\xi_n} = 0$. Since all functions in the integrand are non-negative and continuous, $p_\beta(w_n|\theta)^{\xi_n} p_\pi(w_n)^{1-\xi_n} = 0$ for all θ for the unique i with $\xi_i \neq \xi_i^*$. Since this holds for every element of X, we must have that $p_\beta(w_n|\theta) = 0$ for all $\xi_n = 0$ and $p_\pi(w_n) = 0$ for all $\xi_n = 1$, proving β and π are disjoint, provided the minor assumption that all words in the vocabulary w_n are observed in the data. In practice all words are observed in the vocabulary because we choose the vocabulary based on the training set.

□

9.3 ELBO (per doc)

Let $\Lambda = \{\alpha, \beta, \eta, \delta, \pi, p\}$. Omitting variational parameters for simplicity:

$$\begin{aligned} \log p(\mathbf{w}, \mathbf{y}|\Lambda) &= \log \int_\theta \sum_z \sum_\xi p(\theta, \mathbf{z}, \boldsymbol{\xi}, \mathbf{w}, \mathbf{y}|\Lambda)d\theta \\ &= \log E_q\left(\frac{p(\theta, \mathbf{z}, \boldsymbol{\xi}, \mathbf{w}, \mathbf{y}|\Lambda)}{q(\theta, \mathbf{z}, \boldsymbol{\xi})}\right) \\ &\geq E_q[\log p(\theta, \mathbf{z}, \boldsymbol{\xi}, \mathbf{w}, \mathbf{y})] - E_q[q(\theta, \mathbf{z}, \boldsymbol{\xi})] \end{aligned}$$

Let $ELBO = E_q[\log p(\theta, \mathbf{z}, \boldsymbol{\xi}, \mathbf{w}, \mathbf{y}|\Lambda)] - E_q[q(\theta, \mathbf{z}, \boldsymbol{\xi})]$

Expanding this:

$$\begin{aligned} ELBO = E_q[\log p(\theta|\alpha)] &+ E_q[\log p(\mathbf{z}|\theta)] + E_q[\log p(y|\theta, \eta, \delta)] \\ &+ E_q[\log p(\boldsymbol{\xi}|p)] + E_q[\log p(\mathbf{w}|\mathbf{z}, \beta, \boldsymbol{\xi}, \pi)] \\ &- E_q[\log q(\theta|\gamma)] - E_q[\log q(\mathbf{z}|\phi)] - E_q[\log q(\boldsymbol{\xi}|\varphi)] \end{aligned}$$

The distributions of each of the variables under the generative model are:

$$\begin{aligned}
&\theta_d \sim \text{Dirichlet}(\alpha)\\
&z_{dn}|\theta_d \sim \text{Categorical}(\theta_d)\\
&\xi_{dn} \sim \text{Bernoulli}(p)\\
&w_{dn}|z_{dn}, \xi_{dn} = 1 \sim \text{categorical}(\beta_{z_{dn}})\\
&w_{dn}|z_{dn}, \xi_{dn} = 0 \sim \text{Categorical}(\pi)\\
&y_d|\theta_d \sim \text{GLM}(\theta; \eta, \delta)
\end{aligned}$$

Under the variational posterior, we use the following distributions:

$$\begin{aligned}
&\theta_d \sim \text{Dirichlet}(\gamma_d)\\
&z_{dn} \sim \text{Categorical}(\phi_{dn})\\
&\xi_{dn} \sim \text{Bernoulli}(\varphi_{w_{dn}})
\end{aligned}$$

This leads to the following ELBO terms:

$$\begin{aligned}
E_q[\log p(\theta|\alpha)] &= \log\Gamma\left(\sum_k \alpha_k\right) - \sum_k \log\Gamma(\alpha_k) + \sum_k (\alpha_k - 1)E_q[\log\theta_k]\\
E_q[\log p(\mathbf{z}|\theta)] &= \sum_n \sum_k \phi_{nk} E_q[\log\theta_k]\\
E_q[\log p(\mathbf{w}|\mathbf{z}, \beta, \boldsymbol{\xi}, \pi)] &= \sum_n \left(\sum_v w_{nv}\varphi_v\right) * \left(\sum_k \sum_v \phi_{nk} w_{nv} \log\beta_{kv}\right) + \left(1 - \left(\sum_v w_{nv}\varphi_v\right)\right)\left(\sum_v w_{nv}\log\pi_v\right)\\
E_q[\log p(\boldsymbol{\xi}|p)] &= \sum_n \left(\sum_v w_{nv}\varphi_v\right)\log p + \left(1 - \left(\sum_v w_{nv}\varphi_v\right)\right)\log(1-p)\\
E_q[q(\theta|\gamma)] &= \log\Gamma\left(\sum_k \gamma_k\right) - \sum_k \log\Gamma(\gamma_k) + \sum_k (\gamma_k - 1)E_q[\log\theta_k]\\
E_q[q(\mathbf{z}|\phi)] &= \sum_n \sum_k \phi_{nk}\log\phi_{nk}\\
E_q[q(\boldsymbol{\xi}|\varphi)] &= \sum_n \left(\sum_v w_{nv}\varphi_v\right)\log\left(\sum_v w_{nv}\varphi_v\right) + \left(1 - \left(\sum_v w_{nv}\varphi_v\right)\right)\log\left(1 - \left(\sum_v w_{nv}\varphi_v\right)\right)\\
E_q[\log p(y|\theta, \eta, \delta)] &= \frac{1}{2}\log 2\pi\delta - \frac{1}{2\delta}\left(y^2 - 2y\eta^\top E_q[\theta] + \eta^\top E_q[\theta\theta^\top]\eta\right)
\end{aligned}$$

Other useful terms:

$$\begin{aligned}
E_q[\log\theta_k] &= \Psi(\gamma_k) - \Psi\left(\sum_{j=1}^K \gamma_j\right)\\
\bar{Z} &:= \frac{\sum_n \xi_n z_n}{\sum_n \xi_n} \in \mathbb{R}^K\\
E_q[\theta] &= \frac{\gamma}{\gamma^\top \mathbf{1}}\\
\gamma_0 &:= \sum_k \gamma_k\\
\tilde{\gamma}_j &:= \frac{\gamma_j}{\sum_k \gamma_k}\\
E_q[\theta\theta^\top]_{ij} &= \frac{\tilde{\gamma}_i(\delta(i,j) - \tilde{\gamma}_j)}{\gamma_0 + 1} + \tilde{\gamma}_i\tilde{\gamma}_j
\end{aligned}$$

9.4 Lower Bounds on the Log Likelihood

Remark that the likelihood for the words of one document can be written as follows:

$$p(\mathbf{w}) = \int_\theta d\theta p(\theta|\alpha) \left\{ \prod_{n=1}^{N} [p * p_\beta(w_n|\theta) + (1-p)p_\pi(w_n)] \right\}$$

We would like to derive a lower bound to the joint log likelihood $p(y, \mathbf{w})$ of one document that resembles the prediction constrained log likelihood since they exhibit similar empirical behavior. Write $p(y, \mathbf{w})$ as $E_\xi[p(y|\mathbf{w}, \xi)p(\mathbf{w}|\xi)]$ and apply Jensen's inequality:

$$\log p(y, \mathbf{w}) \geq E_\xi[\log p(y|\mathbf{w}, \xi)] + E_\xi[\log p(\mathbf{w}|\xi)]$$

Focusing on the second term we have:

$$\log p(\mathbf{w}|\xi) = \log \int_\theta d\theta p(\theta|\alpha) \prod_{n=1}^{N} p_\beta(w_n|\theta)^{\xi_n} p_\pi(w_n)^{1-\xi_n}$$

Applying Jensen's inequality again to push the log further inside the integrals:

$$\log p(\mathbf{w}|\xi) \geq \int_\theta d\theta p(\theta|\alpha) \left\{ \sum_{i=1}^{N} \xi_n \log p_\beta(w_n|\theta) + \sum_{n=1}^{N} (1-\xi_n) \log p_\pi(w_n) \right\}$$

Note that θ and ξ are independent, so we have:

$$\log p(y, \mathbf{w}) \geq E[\log p(y|\mathbf{w}, \xi)] + E\left[\sum_{i=1}^{N} \xi_n \log p_\beta(w_n|\theta) + \sum_{n=1}^{N} (1-\xi_n) \log p_\pi(w_n) \right]$$

where the expectation is taken over the ξ and θ priors. This gives the final bound:

$$\log p(y, \mathbf{w}) \geq E[\log p_\beta(y|W_1(\xi))] + pE[\log p_\beta(\mathbf{w}|\theta)] + (1-p) \log p_\pi(\mathbf{w})$$

We have used the substitution: $p(y|\mathbf{w}, \xi) = p_\beta(y|W_1(\xi))$. Conditioning on ξ, y is independent from the set of w_n with $\xi_n = 0$, so we denote $W_1(\xi)$ as the set of w_n with $\xi_n = 1$. It is also clear that $p(y|W_1(\xi), \xi) = p_\beta(y|W_1(\xi))$. By linearity of expectation, this bound can easily be extended to all documents.

Note that this bound is undefined on the constrained parameter space: $\beta^\top \pi = 0$; if $p \neq 0$ and $p \neq 1$. This is clear because $\log p_\pi(\mathbf{w})$ or $\log p_\beta(w_n|\theta)$ is undefined with probability 1. We can also see this directly, since $p(y, \mathbf{w}|\xi)$ is non-zero for exactly one value of ξ so $E[\log p(y, \mathbf{w}|\xi)]$ is clearly undefined. We derive a tighter bound for this particular case as follows. Define $\xi^*(\pi, \beta, \mathbf{w})$ as the unique ξ such that $p(\mathbf{w}|\xi)$ is non-zero. We can write $p(y, \mathbf{w}) = p(y, W|\xi^*(\pi, \beta, \mathbf{w}))p(\xi^*(\pi, \beta, \mathbf{w}))$. For simplicity, I use the notation ξ^* but keep in mind that it's value is determined by β, π and $\mathbf{w}$. Also remark that the posterior of ξ is a point mass as ξ^*. If we repeat the analysis above we get the bound:

$$\log p(y, \mathbf{w}) \geq p_\beta(y|W_1(\xi^*)) + E\left[\sum_{n=1}^{N} \xi^* p_\beta(w_n|\theta) \right] + \sum_{n=1}^{N} (1-\xi^*) \log p_\pi(w_n) + p(\xi^*)$$

which is to be optimized over β and π. Note that the $p(\xi^*)$ term is necessary because of its dependence on β and π. Comparing this objective to our ELBO, we make a number of points. The true posterior is ξ^* which would ordinarily require a combinatorial optimization to estimate; however we introduce the continuous variational approximation $\xi \sim Bern(\varphi)$. Note that the true posterior is a special case of our variational posterior (when $\varphi = 1$ or $\varphi = 0$). Since the parameterization is differentiable, it allows us to estimate ξ^* via gradient descent. Moreover, the parameterization encourages β and π to be disjoint without explicitly searching over the constrained space. Empirically, the estimated set of φ are correct in simulations, and correct given the learned β and π on real data examples.

9.5 Implementation details

In general, we treat α (the prior for the document topic distribution) as fixed (to a vector of ones). We tune pc-SLDA using Hughes et al. [2017b]'s code base, which does a small grid search over relevant parameters. We tune sLDA and pf-sLDA using our own implementation and SGD. β and π are initialized with small, random (exponential) noise to break symmetry. We optimize using ADAM with initial step size 0.025.

We model real targets as coming from $N(\eta^\top \theta, \delta)$ and binary targets as coming from $\text{Bern}(\sigma(\eta^\top \theta))$

9.6 pf-sLDA likelihood and prediction constrained training.

The pf-sLDA marginal likelihood for one document and target can be written as:

$$\begin{aligned} p(\mathbf{w}, y) &= p(y|\mathbf{w}) \int_\theta \sum_\xi p(\mathbf{w}, \theta, \xi) \\ &= p(y|\mathbf{w}) \int_\theta p(\theta|\alpha) \prod_n \{p * p_\beta(w_n|\theta, \xi_n = 1, \beta) + (1-p) * p_\pi(w_n|\xi_n = 0, \pi)\} \end{aligned}$$

where n indexes over the words in the document. We see there still exist the $p(y|\mathbf{w})$ and $p * p_\beta(\mathbf{w})$ that are analagous to the prediction constrained objective, though the precise form is not as clear.

9.7 Data set details

- Pang and Lee's movie review data set [Pang and Lee, 2005]: There are 5006 documents. Each document represents a movie review, and the documents are stored as bag of words and split into 3754/626/626 for train/val/test. After removing stop words and words appearing in more than 50% of the reviews or less than 10 reviews, we get $|V| = 4596$. The target is an integer rating from 1 (worst) to 10 (best).

- Yelp business reviews [Yelp, 2019]: We use a subset of 10,000 documents from the Yelp 2019 Data set challenge . Each document represents a business review, and the documents are stored as bag of words and split into 7500/1250/1250 for train/val/test. After removing stop words and words appearing in more than 50% of the reviews or less than 10 reviews, we get $|V| = 4142$. The target is an integer star rating from 1 to 5.

- Electronic health records (EHR) data set of patients with Autism Spectrum Disorder (ASD), introduced in Masood and Doshi-Velez [2018]: There are 3804 documents. Each document represents the EHR of one patient, and the features are possible diagnoses. The documents are split into 3423/381 for train/val, with $|V| = 3600$. The target is a binary indicator of presence of epilepsy.

- Synthetic: We simulate a set of 5 data sets. Each data set is generated based on the pf-sLDA generative process with β and π random, each with mass on 50 features. This means there are 100 total features, 50 relevant and 50 irrelevant. Other relevant parameters are $K = 5$, $\alpha = \mathbf{1}$, and $p = 0.25$. We use these data sets to test the effectiveness and reliability of the feature filtering of pf-sLDA under the well-specified case when we have ground-truth.

9.8 Full Pang and Lee Movie Review Topics

See Table below.

	sLDA	pc-sLDA, $\lambda = 10$	pf-sLDA, $p = 0.10$
1	motion, way, love, performance, best, picture, films, character, characters, life	best, little, time, good, don, picture, year, rated, films just	wars, emotionally, allows, academy, perspective, tragedy, today, important, oscar, powerful
η_1	7.801	8.287	10.253
2	kind, poor, enjoyable, picture, excellent, money, look, don films, year	little, just, good, doesn, life, way, films, character, time, characters	complex, study, emotions, rare, perfectly, wonderful, unique, power, fascinating, perfect
η_2	5.994	8.127	8.792
3	running, subject, 20, character, message, characters, minutes just, make, time	country, king, stone, dark, political, parker, mood, modern, dance, noir	jokes, idea, wasn, silly, predictable, acceptable, unfortunately, tries, nice, problem
η_3	5.735	4.407	6.258
4	acceptable, language, teenagers, does, make, good, sex , violence, rated, just	subscribe, room, jane, disappointment, michel, screening, primarily, reply, frustrating, plenty	tedious, poorly, horror, dull, acceptable, parody, worse, ridiculous, supposed, bad
η_4	5.064	3.440	2.657
5	plot, time, bad, funny, good, humor, little, isn, action	script, year, little, good, don, look, rated, picture, just, films	suppose, lame, annoying, attempts, failed, attempt, boring, awful, dumb, flat, comedy
η_5	3.516	2.802	0.135

Table 4: We list the top 10 most likely words for each topic for the models specified. The topics are organized from highest to lowest with respect to it's corresponding coefficient for the supervised task. In the context of movie reviews, positive-sentiment words are listed in green, negative-sentiment words are listed in red, and sometimes positive, sometimes negative or neutral, but sentiment-related words are listed in yellow. Non-sentiment related words are listed in black.

9.9 Coherence details

We calculate coherence for each topic by taking the top N most likely words for the topic, calculating the pointwise mutual information for each possible pair, and averaging. These terms are defined below.

$$\begin{aligned}
\text{coherence} &= \frac{1}{N(N-1)} \sum_{w_i, w_j \in \text{TopN}} \text{pmi}(w_i, w_j) \\
\text{pmi}(w_i, w_j) &= \log \frac{p(w_i)p(w_j)}{p(w_i, w_j)} \\
p(w_i) &= \frac{\sum_d I(w_i \in \text{doc d})}{M} \\
p(w_i, w_j) &= \frac{\sum_d I(w_i \text{ and } w_j \in \text{doc d})}{M}
\end{aligned}$$

where M is the total number of documents and N is the number of top words in a topic. The final coherence we report for a model is the average of all the topic coherences.

The numbers in the paper are reported with $N = 50$, but we found that the general trends (i.e. pf-sLDA topics most coherent) were consistent across several values of N. For example, for the Pang and Lee Movie Review Dataset, we have the following coherences:

N	10	20	30	40	50
pc-sLDA	0.42	0.44	0.44	0.46	0.49
pf-sLDA	1.52	1.69	1.76	1.92	1.99

9.10 Coordinate Ascent Updates

9.10.1 E Step

Denote L_φ the terms of the ELBO that depend on φ and similarly for other parameters.

φ update

$$\begin{aligned} L_\varphi = &\sum_n \left(\sum_v w_{nv}\varphi_v\right)\left(\sum_k\sum_v \phi_{nk} w_{nv}\log\beta_{kv}\right) + \left(1-\left(\sum_v w_{nv}\varphi_v\right)\right)\left(\sum_v w_{nv}\log\pi_v\right) \\ &+\sum_n \left(\sum_v w_{nv}\varphi_v\right)\log p + \left(1-\left(\sum_v w_{nv}\varphi_v\right)\right)\log(1-p) \\ &-\sum_n \left(\sum_v w_{nv}\varphi_v\right)\log\left(\sum_v w_{nv}\varphi_v\right) + \left(1-\left(\sum_v w_{nv}\varphi_v\right)\right)\log\left(1-\left(\sum_v w_{nv}\varphi_v\right)\right) \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial\varphi_j} L_\varphi = &\sum_n w_{nj}\left(\sum_k\sum_v \phi_{nk} w_{nv}\log\beta_{kv}\right) - w_{nj}\left(\sum_v w_{nv}\log\pi_v\right) \\ &+\sum_n w_{nj}(\log p - \log(1-p)) \\ &-\sum_n w_{nj}\left(\log\left(\sum_v w_{nv}\varphi_v\right) - \log\left(1-\left(\sum_v w_{nv}\varphi_v\right)\right)\right) \\ =&\sum_n w_{nj}\left(\sum_k \phi_{nk}\log\beta_{kj} - \log\pi_j\right) \\ &+\sum_n w_{nj}(\log p - \log(1-p)) \\ &-\sum_n w_{nj}\left(\log\varphi_j - \log\left(1-\varphi_j\right)\right) \end{aligned}$$

Let

$$\Omega_j = \sum_n w_{nj}\left(\sum_k \phi_{nk}\log\beta_{kj} - \log\pi_j + \log p - \log(1-p)\right)$$

.

Setting the gradient to 0 and solving, we get the following update:

$$\varphi_v \leftarrow \frac{\exp(\Omega_j/\sum_n w_{nj})}{1+\exp(\Omega_j/\sum_n w_{nj})}$$

This update makes intuitive sense. It looks like the ratio between the probability of β and π explaining the word, weighted by p, normalized, sigmoided to make it a valid probability.

ϕ update

Assuming y depends on θ for now.

$$
\begin{aligned}
L_\phi =& \sum_n \sum_k \phi_{nk} \left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^K \gamma_j \right) \right) \\
&+ \sum_n \left(\sum_v w_{nv} \varphi_v \right) \left(\sum_k \sum_v \phi_{nk} w_{nv} \log \beta_{kv} \right) \\
&- \sum_n \sum_k \phi_{nk} \log \phi_{nk}
\end{aligned}
$$

Let $w_n = v$ and using Lagrange Multipliers with the constraint $\sum_k \phi_{nk} = 1$:

$$
\frac{\partial}{\partial \phi_{nk}} L_\phi = \Psi(\gamma_k) - \Psi\left(\sum_{j=1}^K \gamma_j \right) + \varphi_v \log \beta_{kv} - 1 - \log \phi_{nk} + \lambda
$$

This gives the update:

$$
\phi_{nk} \propto \beta_{kv}^{\varphi_v} \exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^K \gamma_j \right) \right)
$$

This is the same update as LDA, with β weighted by φ.

γ Update

$$
\begin{aligned}
L_\gamma =& \sum_k (\alpha_k - 1) \left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^K \gamma_j \right) \right) \\
&+ \sum_n \sum_k \phi_{nk} \left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^K \gamma_j \right) \right) \\
&- \log \Gamma\left(\sum_k \gamma_k \right) - \sum_k \log \Gamma(\gamma_k) + \sum_k (\gamma_k - 1) \left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^K \gamma_j \right) \right) \\
&- \frac{1}{2\sigma^2} \left(y^2 - 2y\eta^\top E_q[\theta] + \eta^\top E_q[\theta\theta^\top]\eta \right)
\end{aligned}
$$

The gradient of γ consists of two components; the first (for the first three lines) is the same as in LDA:

$$
\frac{\partial}{\partial \gamma_i} L1_\gamma = \Psi'(\gamma_i) \left(\alpha_i + \sum_n \phi_{ni} - \gamma_i \right) - \Psi'\left(\sum_k \gamma_k \right) \sum_k \left(\alpha_k + \sum_n \phi_{nk} - \gamma_k \right)
$$

The second term is as follows:

$$\begin{aligned}\frac{\partial L2_\gamma}{\partial \gamma_i} =& -\frac{1}{2\sigma^2}[-2y_d(\frac{\eta_k}{\gamma_0} - \frac{\sum_{j=1}^K \eta_i \gamma_{dj}}{\gamma_0^2}) + 2\sum_{i \neq k} \eta_i \eta_k \frac{-\gamma_{di}\gamma_0^2(\gamma_0+1) + \gamma_{dk}\gamma_{di}(2\gamma_0(\gamma_0+1)+\gamma_0^2)}{\gamma_0^4(\gamma_0+1)^2} \\ &+ \eta_k^2 \frac{(1-2\gamma_{dk})\gamma_0^2(\gamma_0+1) + \gamma_{dk}(\gamma_{dk}-1)(2\gamma_0(\gamma_0+1)+\gamma_0^2)}{\gamma_0^4(\gamma_0+1)^2} + 2\sum_{i \neq k} \eta_k \eta_i \frac{\gamma_i \gamma_0^2 - 2\gamma_0 \gamma_i \gamma_k}{\gamma_0^4} \\ &+ \eta_k^2 \frac{2\gamma_k \gamma_0^2 - 2\gamma_0 \gamma_k^2}{\gamma_0^4}]\end{aligned}$$

9.10.2 M Step

π Update

$$L_\pi = \sum_n \left(1 - \left(\sum_v w_{nv}\varphi_v\right)\right)\left(\sum_v w_{nv}\log \pi_v\right)$$

Taking the derivative and using lagrange multipliers with the constraint $\sum_v \pi_v = 1$, we get:

$$\begin{aligned}\frac{\partial}{\partial \pi_j} L_\pi &= \sum_n \left(1 - \sum_v w_{nv}\varphi_v\right)\frac{w_{nj}}{\pi_j} + \lambda \\ &= \sum_n (1-\varphi_j)\frac{w_{nj}}{\pi_j} + \lambda\end{aligned}$$

This gives the update:

$$\pi_j \propto (1-\varphi_j)\sum_n w_{nj}$$

This is an intuitive update - it is the proportion the word appears, weighted by φ.

β update

$$L_\beta = \sum_n \left(\sum_v w_{nv}\varphi_v\right)\left(\sum_k \sum_v \phi_{nk} w_{nv} \log \beta_{kv}\right)$$

Adding a Lagrange multiplier for constraint $\sum_v \beta_{kv} = 1 \; \forall k$ and then taking gradient:

$$\frac{\partial}{\partial \beta_{kj}} L_\beta = \sum_n \frac{\varphi_j \phi_{nk} w_{nj}}{\beta_{kv}} + \lambda$$

This gives the update:

$$\beta_{kj} \propto \sum_n \varphi_j \phi_{nk} w_{nj}$$

This is the same update as sLDA, but weighted by φ.

α update

This should be exactly the same as in LDA and sLDA via Newton-Raphson.

η update

Same form as sLDA updates, but replace $E[\bar{Z}]$ and $E[\bar{Z}\bar{Z}^\top]$ with $E[\theta]$ and $E[\theta\theta^\top]$.

δ **update**

Same form as sLDA updates, but replace $E[\bar{Z}]$ and $E[\bar{Z}\bar{Z}^\top]$ with $E[\theta]$ and $E[\theta\theta^\top]$.