Prediction Focused Topic Models Via Vocab Filtering

Russell Kunes * rk3064@columbia.edu
Jason Ren * jason_ren@college.harvard.edu
Finale Doshi-Velez finale@seas.harvard.edu

Abstract

Supervised topic models are often sought to balance prediction quality and interpretability. However, when models are (inevitably) misspecified, standard approaches rarely deliver on both. We introduce a novel approach, the prediction-focused topic model, that uses the supervisory signal to retain only vocabulary terms that improve, or do not hinder, prediction performance. By removing terms with irrelevant signal, the topic model is able to learn task-relevant, interpretable topics. We demonstrate on several data sets that compared to existing approaches, prediction-focused topic models are able to learn much more coherent topics while maintaining competitive predictions.

1 Introduction

Supervised topic models are often sought to balance prediction quality and interpretability (i.e. Hughes et al. [2017a], Kuang et al. [2017]). However, standard supervised topic models often learn topics that are not discriminative in target space. Even in the best of cases, these methods must explicitly trade-off between predicting the target and explaining the count data well [Hughes et al., 2017b]. In this work, we focus on one common reason why supervised topic models fail: documents often contain terms with high occurrence that are irrelevant to the task. For example, in sentiment analysis on movie reviews, topics assign high probability to words like "comedy", "action", "character", and "plot," which may be nearly orthogonal to the sentiment label. The existence of features irrelevant to the supervised task complicates optimization of the trade-off between prediction quality and explaining the count data, and also renders the topics less interpretable.

To address this issue, we introduce a novel supervised topic model, prediction-focused sLDA (pf-sLDA), that explicitly severs the connection between irrelevant features and the response variable and a corresponding variational inference procedure that enforces our parameter constraints. We demonstrate that pf-sLDA outperforms existing approaches with respect to topic coherence on several data sets, while maintaining competitive prediction quality. The full version of this extended abstract can be viewed at: `https://arxiv.org/pdf/1910.05495.pdf`.

2 Related Work

Improving prediction quality in supervised topic models. Since the original supervised LDA (sLDA) work of Mcauliffe and Blei [2008], many works have incorporated the prediction target into the topic model training process in different ways to improve prediction quality, including power-sLDA [Zhang and Kjellström, 2014], med-LDA [Zhu et al., 2012], BP-sLDA [Chen et al., 2015]. Hughes et al. [2017b] pointed out a number of shortcomings of these previous methods and introduced a new objective that weights a combination of the conditional likelihood and marginal data likelihood: $\lambda \log p(\mathbf{y}|\mathbf{w}) + \log p(\mathbf{w})$. They demonstrated the resulting method, termed prediction-constrained sLDA (pc-sLDA), achieves better empirical results in optimizing the trade-off between prediction

*equal contribution

Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. (non-archival)

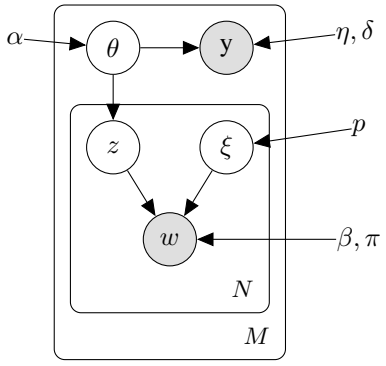
quality and explaining the count data and justify why this is the case. However, their topics are often polluted by irrelevant terms. The pf-sLDA formulation enjoys analogous theoretical properties but effectively removes irrelevant terms, and thus achieves more coherent topics.

Focusing learned topics. The notion of focusing topics in relevant directions is also present in the unsupervised topic modeling literature. For example, Wang et al. [2016] focus topics by seeding them with keywords; Kim et al. [2012] introduce variable selection for LDA, which models some of the vocabulary as irrelevant. Fan et al. [2017] similarly develop stop-word exclusion schemes. However, these approaches adjust topics based on some general notions of "focus", whereas pf-sLDA removes irrelevant signal for a supervised task to explicitly manage a trade-off between prediction quality and explaining the count data.

3 Background and Notation

We briefly describe supervised Latent Dirichlet Allocation (sLDA) [Mcauliffe and Blei, 2008], which our work builds off. sLDA models count data (words) as coming from a mixture of K topics $\{\beta_k\}_{k=1}^K$, where each topic $\beta_k \in \Delta^{|V|-1}$ is a categorical distribution over a vocabulary V of $|V|$ discrete features (words). The count data are represented as a collection of M documents, with each document $\mathbf{w}_d \in \mathbb{N}^{|V|}$ being a vector of counts over the vocabulary. Each document d is associated with a target y_d. Additionally, each document has an associated topic distribution $\theta_d \in \Delta^{K-1}$, which generates both the words and the target.

4 Prediction Focused Topic Models



for each document **do**
  Draw topic distribution $\theta \sim Dir(\alpha)$
  for each word **do**
    Draw topic $z \sim Cat(\theta)$
    Draw switch $\xi \sim Bern(p)$
    if $\xi == 1$ **then**
      Draw word $w \sim Cat(\beta_z)$
    else if $\xi == 0$ **then**
      Draw word $w \sim Cat(\pi)$
  Draw target $y \sim GLM(\theta; \eta, \delta)$
  (See Appendix 8.4 for details)

Figure 1: *Left*: pf-sLDA graphical model. *Right*: pf-sLDA generative process per document

We now introduce prediction-focused sLDA (pf-sLDA). The fundamental assumption that pf-sLDA builds on is that the vocabulary V can be divided into two disjoint components, one of which is irrelevant to predicting the target variable. pf-sLDA separates out the words irrelevant to predicting the target, even if they have latent structure, so that the topics can focus on only modelling structure that is relevant to predicting the target.

Generative Model. The pf-sLDA latent variable model has the following components: one channel of pf-sLDA models the count data as coming from a mixture of K topics $\{\beta_k\}_{k=1}^K$, similar to sLDA. The second channel of pf-sLDA models the data as coming from an additional topic $\pi \in \Delta^{|V|-1}$. The target only depends on the first channel, so the second channel acts as an outlet for words irrelevant to predicting the target. We constrain β and π such that $\beta_k^\top \pi = 0$ for all k, such that each word is always either relevant or irrelevant to predicting the target. Which channel a word comes from is determined by its corresponding Bernoulli switch, which has prior p. The generative process of pf-sLDA is given in Figure 1. In Appendix 8.3, we prove that a lower bound to the pf-sLDA log likelihood is:

$$\log p(\mathbf{y}, \mathbf{w}) \geq E_{\boldsymbol{\xi}}[\log p_\beta(\mathbf{y}|\mathbf{w}, \boldsymbol{\xi})] + pE_{\boldsymbol{\theta}}[\log p_\beta(\mathbf{w}|\boldsymbol{\theta})] + (1-p)\log p_\pi(\mathbf{w}) \tag{1}$$

Connection to prediction-constrained models. The lower bound above reveals a connection to the pc-sLDA loss function. Similar connections can be seen in the true likelihood as described in

Appendix 8.5, but we use the bound for clarity. The first two terms capture the trade-off between performing the prediction task $E_{\boldsymbol{\xi}}[\log p_\beta(\mathbf{y}|\mathbf{w}, \boldsymbol{\xi})]$ and explaining the words $pE_{\boldsymbol{\theta}}[\log p_\beta(\mathbf{w}|\boldsymbol{\theta})]$, where the switch prior p is used to down-weight the latter task (or emphasize the prediction task). This is analogous to the prediction-constrained objective, but we manage the trade-off through an interpretable model parameter, the switch prior p, rather than a more arbitrary Lagrange multiplier λ.

5 Inference

Inference in the pf-sLDA framework corresponds to inference in a graphical model, so advances in Bayesian inference can be applied to solve the inference problem. In this work, we take a variational approach. Our objective is to maximize the evidence lower bound (full form specified in Appendix 8.2), with the constraint that the relevant topics β and additional topics π have disjoint support. The key difficulty is that of optimizing over the non-convex set $\{\beta, \pi : \beta^\top \pi = \mathbf{0}\}$. We resolve this with a strategic choice of variational family, which results in a straightforward training procedure that does not require any tuning parameters.

$$q(\theta, \mathbf{z}, \boldsymbol{\xi}|\phi, \varphi, \gamma) = \prod_d q(\theta_d|\gamma_d) \prod_n q(\xi_{dn}|\varphi) q(z_{dn}|\phi_{dn})$$

$$\theta_d|\gamma_d \sim \text{Dir}(\gamma_d),\ z_{dn}|\phi_{dn} \sim \text{Cat}(\phi_{dn}),\ \xi_{dn}|\varphi \sim \text{Bern}(\varphi_{w_{dn}})$$

The proof of why this variational family enforces our desired constraint is given in Appendix 8.1. To train, we run stochastic gradient descent on the evidence lower bound (ELBO).

6 Experimental Results

6.1 Experimental Set-Up

Metrics. We wish to assess prediction quality and interpretability of learned topics. To measure prediction quality, we use RMSE for real targets and AUC for binary targets. To measure interpretability of topics, we use normalized pointwise mutual information coherence, which was shown by Newman et al. [2010] to be the metric that most consistently and closely matches human judgement in evaluating interpretability of topics. See Appendix 8.6 for coherence calculation details.

Baselines. The recent work in Hughes et al. [2017b] demonstrates that pc-sLDA outperforms other supervised topic modeling approaches, so we use pc-sLDA as our main baseline. We also include standard sLDA [Mcauliffe and Blei, 2008] for reference.

Data Sets. We run our model and baselines on three data sets (see Appendix 8.7 for details):

- Pang and Lee's movie review data set [Pang and Lee, 2005]: 5006 movie reviews, with integer ratings from 1 (worst) to 10 (best) as targets.
- Yelp business reviews [Yelp, 2019]: 10,000 business reviews, with integer stars from 1 (worst) to 5 (best) as targets.
- Electronic health records (EHR) of patients with Autism Spectrum Disorder (ASD) [Masood and Doshi-Velez, 2018]: 3804 EHRs, with binary indicator of epilepsy as target.

Implementation details. Refer to Appendix 8.4

6.2 Results

pf-sLDA learns the most coherent topics. Across data sets, pf-sLDA learns the most coherent topics by far (see Figure 2). pc-sLDA improves on topic coherence compared to sLDA, but cannot match the performance of pf-sLDA. Qualitative examination of the topics in Table 2 supports the claim that the pf-sLDA topics are more coherent, more interpretable, and more focused on the supervised task.

Prediction quality of pf-sLDA remains competitive. pf-sLDA produces similar prediction quality compared to pc-sLDA across data sets (see Figure 2). Both pc-sLDA and pf-sLDA outperform sLDA in prediction quality. In the best performing models of pf-sLDA for all 3 data sets, generally between 10% and 20% of the words were considered relevant. Considering both more words or less words relevant hurt performance, as seen in the plots in Figure 2.

Pang and Lee Movie Reviews		
Model	Coherence	RMSE
sLDA	0.362 (0.101)	1.682 (0.021)
pc-sLDA	1.296 (0.130)	**1.298** (0.015)
pf-sLDA	**2.810** (0.092)	1.305 (0.024)

Yelp Reviews		
Model	Coherence	RMSE
sLDA	0.848 (0.086)	1.162 (0.017)
pc-sLDA	1.080 (0.213)	0.953 (0.004)
pf-sLDA	**3.258** (0.102)	**0.952** (0.011)

ASD Dataset		
Model	Coherence	AUC
sLDA	1.412 (0.113)	0.590 (0.013)
pc-sLDA	2.178 (0.141)	0.701 (0.015)
pf-sLDA	**2.639** (0.091)	**0.748** (0.013)

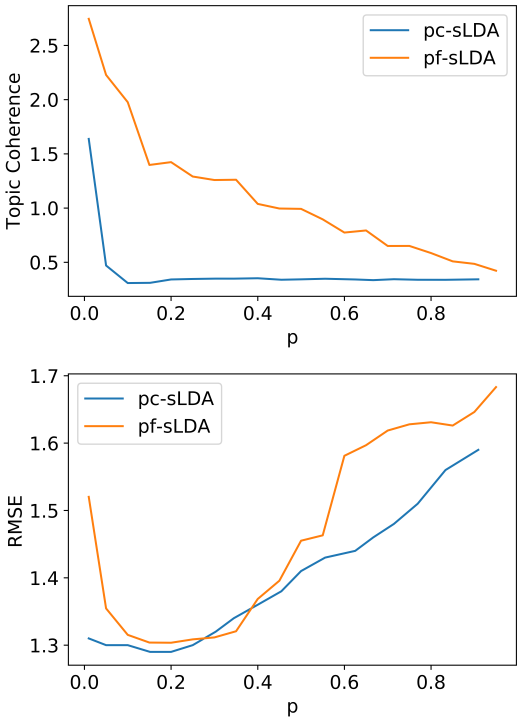


Figure 2: *Left* Mean and (SD) across 5 runs for topic coherence (higher is better) and RMSE (lower is better) or AUC (higher is better) on held-out test sets. Final models were chosen based on a combination of validation coherence and RMSE/AUC. pf-sLDA produces topics with much higher coherence across all three data sets, while maintaining similar prediction performance. *Right:* Validation coherence and RMSE on Pang and Lee Movie Review data set as $p = \frac{1}{\lambda}$ varies. This demonstrates the effect of the channel switch prior p controlling the trade-off between prediction quality and explaining the count data.

Pang and Lee Movie Reviews			
	sLDA	pc-sLDA	pf-sLDA
High	motion, way, love, performance, best, picture, films, character, characters, life	best, little, time, good, don, picture, year, rated, films just	brilliant, rare, perfectly, true, oscar, documentary, wonderful, fascinating, perfect, best
Low	plot, time, bad, funny, good, humor, little, isn, action	script, year, little, good, don, look, rated, picture, just, films	awful, stupid, gags, dumb, dull, sequel, flat, worse, ridiculous, bad

Table 1: We list the most probable words in the topics with the highest and lowest regression coefficient for each model for Pang and Lee Movie Reivews (See Appendix 8.8 for topics for other data sets). Words expected to be in a high regression coefficient topic are listed in green, and words expected to be in a low regression coefficient topic are listed in red. It is clear that the topics learned by pf-sLDA are the most coherent and contain the most words with task relevance.

7 Conclusion

In this paper, we introduced prediction-focused supervised LDA, whose vocabulary selection procedure improves topic coherence of supervised topic models while maintaining competitive prediction quality. The model enjoys good theoretical properties, inferential properties, and produced good empirical results. Future work could include establishing additional theoretical properties of the pf-sLDA variable selection procedure, and applying our trick of managing trade-offs within a graphical model for variable selection in other generative models.

References

Jianshu Chen, Ji He, Yelong Shen, Lin Xiao, Xiaodong He, Jianfeng Gao, Xinying Song, and Li Deng. End-to-end learning of lda by mirror-descent back propagation over a deep architecture. In *Advances in Neural Information Processing Systems*, pages 1765–1773, 2015.

Angela Fan, Finale Doshi-Velez, and Luke Miratrix. Prior matters: simple and general methods for evaluating and improving topic quality in topic modeling. 2017.

Michael C Hughes, Gabriel Hope, Leah Weiner, Thomas H McCoy, Roy H Perlis, Erik B Sudderth, and Finale Doshi-Velez. Prediction-constrained topic models for antidepressant recommendation. *arXiv preprint arXiv:1712.00499*, 2017a.

Michael C Hughes, Leah Weiner, Gabriel Hope, Thomas H McCoy Jr, Roy H Perlis, Erik B Sudderth, and Finale Doshi-Velez. Prediction-constrained training for semi-supervised mixture and topic models. *arXiv preprint arXiv:1707.07341*, 2017b.

Dongwoo Kim, Yeonseung Chung, and Alice H. Oh. Variable selection for latent dirichlet allocation. *ArXiv*, abs/1205.1053, 2012.

Da Kuang, P. Jeffrey Brantingham, and Andrea L. Bertozzi. Crime topic modeling. *Crime Science*, 6: 1–20, 2017.

M. Arjumand Masood and Finale Doshi-Velez. A particle-based variational approach to bayesian non-negative matrix factorization. *J. Mach. Learn. Res.*, 20:90:1–90:56, 2018.

Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *HLT-NAACL*, 2010.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*, 2005.

Shuai Wang, Zhiyuan Chen, Geli Fei, Bing Liu, and Sherry Emery. Targeted topic modeling for focused analysis. In *KDD*, 2016.

Yelp. Yelp dataset challenge, 2019. data retrieved from `https://www.yelp.com/dataset/challenge`.

Cheng Zhang and Hedvig Kjellström. How to supervise topic models. In *European Conference on Computer Vision*, pages 500–515. Springer, 2014.

Jun Zhu, Amr Ahmed, and Eric P Xing. Medlda: maximum margin supervised topic models. *Journal of Machine Learning Research*, 13(Aug):2237–2278, 2012.

8 Appendix

8.1 Variational Family

We show how our choice of variational family incorporates our desired constraint in the model parameters. The constraint we wish to satisfy is $\beta^\top \pi = \mathbf{0}$. Our choice of variational family is:

$$\begin{aligned} q(\theta, \mathbf{z}, \boldsymbol{\xi}|\phi, \varphi, \gamma) &= \prod_d q(\theta_d|\gamma_d) \prod_n q(\xi_{dn}|\varphi) q(z_{dn}|\phi_{dn}) \\ \theta_d|\gamma_d &\sim \text{Dir}(\gamma_d) \\ z_{dn}|\phi_{dn} &\sim \text{Cat}(\phi_{dn}) \\ \xi_{dn}|\varphi &\sim \text{Bern}(\varphi_{w_{dn}}) \end{aligned}$$

where d indexes over the documents and n indexes over the words in each document. We first propose two theorems relating to the model.

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 ξ^* s.t. $p(\xi^*|\mathbf{w}, y) = 1$*

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 $p(\xi^*|\mathbf{w}, y) = 1$ for the ξ^* corresponding to $\mathbf{w}$ as described before.

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

   Then we have:

$$p(\xi^*|\mathbf{w}, y) = \frac{p(\mathbf{w}, y|\xi^*)p(\xi^*)}{\sum_\xi p(\mathbf{w}, y|\xi)p(\xi)} = 1$$
$$p(\mathbf{w}, y|\xi^*)p(\xi^*) = \sum_\xi p(\mathbf{w}, y|\xi)p(\xi)$$

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

   Then we have:

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

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

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

   Let X be the set of ξ that differ from ξ^* in one and only one position, i.e. $\xi_n = \xi_n^*$ for all $n \in \{1, \dots N\} \setminus \{i\}$ and $\xi_i \neq \xi_i^*$. For each $\xi \in X$, $\int_\theta p(\theta)\prod_n p_\beta(w_n|\theta)^{\xi_n} p_\pi(w_n)^{1-\xi_n} = 0$. 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 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.

□

Theorems 1 and 2 tell us that if the posterior distribution of the channel switches $\boldsymbol{\xi}_d$ is independent of the posterior distribution of the document topic distribution θ_d for all documents, then the true relevant topics β and additional topic π must have disjoint support, and moreover the posterior of the channel switches $\boldsymbol{\xi}$ is a point mass. This suggests that to enforce that β and π are disjoint, we should choose the variational family such that $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are independent.

If $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are conditionally independent in the posterior, then the posterior can factor as $p(\boldsymbol{\xi}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{w}) = p(\boldsymbol{\xi}|\mathbf{y}, \mathbf{w})p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{w})$. In this case, the posterior for the channel switch of the nth word in document d, ξ_{dn}, has no dependence d, which can be seen directly from the graphical model. Thus, choosing $q(\boldsymbol{\xi}|\varphi)$ to have no dependence on document naturally pushes our assumption into the variational posterior.

Our choices for the variational distributions for $\boldsymbol{\theta}$ and $\mathbf{z}$ match those of Mcauliffe and Blei [2008]. We choose $q(\xi_{dn}|\varphi_{w_{dn}})$ to be a Bernoulli probability mass function with parameter $\varphi_{w_{dn}}$ indexed only by the word w_{dn}. This distribution acts as a relaxation of a true point mass posterior, allowing us to use gradient information to optimize over $[0, 1]$ rather than directly over $\{0, 1\}$. Moreover, this parameterization allows us to naturally use the variational parameter φ as a feature selector; low estimated values of φ indicate irrelevant words, while high values of φ indicate relevant words.

8.2 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:

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

Other useful terms:

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

8.3 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.

8.4 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. Our code base will be made public in the near future. 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 Bern$(\sigma(\eta^\top \theta))$

8.5 pf-sLDA likelihood and prediction constrained training.

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

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

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 analogous to the prediction constrained objective, though the precise form is not as clear.

8.6 Coherence details

We calculate coherence for each topic by taking the top 50 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 $N = 50$ 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.

8.7 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.

8.8 Qualitative Topic Examination

Pang and Lee Movie Reviews			
	sLDA	pc-sLDA	pf-sLDA
High	motion, way, love, performance, best, picture, films, character, characters, life	best, little, time, good, don, picture, year, rated, films just	brilliant, rare, perfectly, true, oscar, documentary, wonderful, fascinating, perfect, best
Low	plot, time, bad, funny, good, humor, little, isn, action	script, year, little, good, don, look, rated, picture, just, films	awful, stupid, gags, dumb, dull, sequel, flat, worse, ridiculous, bad
Yelp Reviews			
	sLDA	pc-sLDA	pf-sLDA
High	fries, fresh, burger, try, cheese, really, pizza, place, like, good	best, just, amazing, love, good, food, service, place, time, great	fantastic, loved, highly, fun, excellent, awesome, atmosphere, amazing, delicious, great
Low	store, time, want, going, place, know, people, don, like, just	didn, don, said, told, like, place, time, just, service, food	awful, management, dirty, poor, horrible, worst, rude, terrible, money, bad
ASD			
	sLDA	pc-sLDA	pf-sLDA
High	Intellect disability Esophageal reflux Hearing loss Development delay Downs syndrome	Infantile cerebral palsy Congenital quadriplegia Esophageal reflux fascia Muscle/ligament dis Feeding problem	Other convulsions Aphasia Convulsions Central hearing loss Grand mal status
Low	Otitis media Asthma Downs syndrome Scoliosis Constipation	Accommodative esotropia Joint pain-ankle Congenital factor VIII Fragile X syndrome Pain in limb	Autistic disorder Diabetes Type 1 c0375114 Other symbolic dysfunc Diabetes Type 1 c0375116 Diabetes Type 2

Table 2: We list the most probable words in the topics with the highest and lowest regression coefficient for each model and data set. In the context of each data set, for ease of evaluation, words expected to be in a high regression coefficient topic are listed in green, and words expected to be in a low regression coefficient topic are listed in red. It is clear that the topics learned by pf-sLDA are the most coherent and contain the most words with task relevance.