Supplement: Semi-Supervised Prediction-Constrained Topic Models

Michael C. Hughes[*1], **Gabriel Hope**[2], **Leah Weiner**[3],
Thomas H. McCoy, Jr.[4], **Roy H. Perlis**[4], **Erik B. Sudderth**[2,3], and **Finale Doshi-Velez**[1]

[1]Harvard University SEAS, *Email: mike@michaelchughes.com
[2]School of Information & Computer Sciences, Univ. of California, Irvine
[3]Dept. of Computer Science, Brown University
[4]Massachusetts General Hospital & Harvard Medical School

Abstract

This document contains supplementary material to the AISTATS 2018 accepted paper "Semi-Supervised Prediction-Constrained Topic Models."

Contents

Further Resources

Public code: `https://github.com/dtak/prediction-constrained-topic-models/`

Earlier workshop paper: Hughes, Hope, Weiner, McCoy, Perlis, Sudderth, and Doshi-Velez. "Prediction-Constrained Topic Models for Antidepressant Prediction." In NIPS 2017 Workshop on Machine Learning for Health (NIPS ML4H 2017). `https://arxiv.org/abs/1712.00499`

Longer tech report: Hughes, Weiner, Hope, McCoy, Perlis, Sudderth, and Doshi-Velez. "Prediction-Constrained Training for Semi-Supervised Mixture and Topic Models." arXiv e-print 2017. `https://arxiv.org/abs/1707.07341`

A Comparison of Supervised Latent Variable Training Frameworks

In this section, we expand on the main paper's discussion of previous frameworks for training supervised latent variable models. In particular, we recap our formal justification for our proposed prediction-constrained (PC) training objective and provide detailed mathematical comparisons to alternative objectives.

Recall that our fundamental contribution is the prediction-constrained training objective:

$$\min_{\phi,\eta} - \Big[\sum_{d=1}^{D} \log p(x_d|\phi,\alpha) \Big] - \log p(\phi,\eta), \tag{1}$$
$$\text{subject to} \quad - \sum_{d=1}^{D} \log p(y_d|x_d,\phi,\eta,\alpha) \leq \epsilon.$$

Lagrange multiplier theory allows us to transform the inequality constrained objective above to an equivalent unconstrained problem:

$$\min_{\phi,\eta} - \sum_{d=1}^{D} \Big[\log p(x_d|\phi) + \lambda_\epsilon \log p(y_d|x_d,\phi,\eta) \Big] \quad - \log p(\phi,\eta). \tag{2}$$

Here $\lambda_\epsilon > 0$ is a scalar Lagrange multiplier. For each distinct value of λ_ϵ, a solution to Eq. (2) matches a solution to Eq. (1) for some ϵ. The relationship between λ_ϵ and ϵ is monotonic, but it does not have a known analytic form; we must search over the one-dimensional space of penalties λ_ϵ for an appropriate value.

We can further expand our unconstrained PC training objective by making the marginalization over the hidden variables π_d explicit:

$$\min_{\phi,\eta} - \sum_{d=1}^{D} \Big[\log p(x_d|\phi) + \log \Big(\int_{\pi_d} p(y_d|\pi_d,\eta) p(\pi_d|x_d,\phi,\alpha) d\pi_d \Big)^{\lambda_\epsilon} \Big] \quad - \log p(\phi,\eta). \tag{3}$$

While the definition of the PC training objective in Eq. (2) is straightforward, it has desirable features that are not shared by other supervised training objectives for topic models. In this appendix we contrast the PC objective with several other approaches, often comparing to methods from the topic modeling literature to give concrete alternatives.

A.1 Advantages over standard joint likelihood training

The most standard training method is to find a point estimate of global parameters ϕ, η that maximizes the (regularized) joint log-likelihood $\log p(x, y \mid \phi, \eta)$. Related Bayesian methods that approximate the posterior distribution $p(\phi, \eta \mid x, y)$, such as variational methods (Wainwright and Jordan, 2008) and Markov chain Monte Carlo methods (Andrieu et al., 2003), estimate moments of the same *joint* likelihood relating hidden variables π_d to data x_d and labels y_d.

For example, supervised LDA (McAuliffe and Blei, 2008; Wang et al., 2009) learns latent topic assignments π_d by optimizing the joint probability of bag-of-words document representations x_d and document labels y_d. One of several problems with this joint likelihood objective is *cardinality mismatch*: the relative sizes of the random variables x_d and y_d can reduce predictive performance. In particular, if y_d is a one-dimensional binary label but x_d is a high-dimensional word count vector, the optimal solution to $\max_{\phi,\eta} \log p(x_d, y_d|\phi,\eta)$ will often be indistinguishable from the solution to the *unsupervised* problem of modeling the data x alone. Low-dimensional labels can have neglible impact on the joint density compared to the high-dimensional words x_d, causing learning to ignore subtle features that are critical for the prediction of y_d from x_d. Despite this issue, recent work continues to use this training objective (Wang and Zhu, 2014; Ren et al., 2017).

A.2 Advantages over maximum conditional likelihood training

Motivated by similar concerns about joint likelihood training, Jebara and Pentland (1999) introduce a method to explicitly optimize the conditional likelihood $\log p(y \mid x, \phi, \eta)$ for the Gaussian mixture model. They replace the conditional likelihood with a more tractable lower bound, and then

monotonically increase this bound via a coordinate ascent algorithm they call *conditional expectation maximization* (CEM). Chen et al. (2015) instead use a variant of backpropagation to optimize the conditional likelihood of a supervised topic model.

One concern about the conditional likelihood objective is that it *exclusively* focuses on the prediction task; it need not lead to good models of the data x, and it cannot incorporate unlabeled data. In contrast, our prediction-constrained (PC) training allows a principled tradeoff between optimizing the marginal likelihood of data and the conditional likelihood of labels given data. Plus, PC training naturally handles partially labeled datasets.

A.3 Advantages over label replication

We are not the first to notice that high-dimensional data x_d can swamp the influence of low-dimensional labels y_d. Among practitioners, one common workaround to this imbalance is to retain the symmetric maximum joint likelihood objective, but to *replicate* each label y_d as if it were observed R times per document: $\{y_d, y_d, \ldots, y_d\}$. Applied to supervised LDA, label replication leads to an alternative *power sLDA* topic model (Zhang and Kjellström, 2014).

Label replication still leads to nearly the same per-document joint density, except that the likelihood density is raised to the R-th power: $p(y_d \mid \pi_d, \eta)^R$. While label replication can better "balance" the relative sizes of x_d and y_d when $R \gg 1$, performance gains over standard supervised LDA are often negligible because this approach does not address the *assymmetry issue*. To see why, we examine the label-replicated training objective when written as an integral over the hidden variable π_d:

$$\min_{\phi,\eta} \quad -\sum_{d=1}^{D} \log \left[\int p(\pi_d \mid \alpha) p(x_d \mid \pi_d, \phi) p(y_d \mid \pi_d, \eta)^R \, d\pi_d \right] - \log p(\phi, \eta). \tag{4}$$

It is worthwhile to contrast the label replication objective here in Eq. (4) with our PC objective in Eq. (3), to see that they are formally distinct. Eq. (4) upweights the label likelihood $p(y_d|\pi_d)$ *inside the integral* over π_d, while our PC approach upweights the *entire integral* $p(y_d|x_d) = \int p(y_d, \pi_d \mid x_d) d\pi_d$. Thus, our PC approach emphasizes the asymmetric task of predicting labels from data (y_d from x_d), while label replication only emphasizes the connection between labels from hiddens (y_d from π_d).

It is easy to find examples where the optimal solution to the label replication objective performs poorly on the target task of predicting y given only x, because the training has not directly prioritized this asymmetric prediction. In the main paper, Fig. 1 provides an intuition-building example where maximum likelihood training with label replication fails to give good prediction performance for *any* value of the replication weight $R > 1$, while our PC approach can do much better when λ is sufficiently large. (Note: In Fig. 1 the replication weight R is renamed as λ). Crucially, the reason our approach is better is that it is more resistant to model misspecification, while label replication requires the model assumptions about both x and y to become *more and more correct* as R increases.

A.4 Advantages over posterior regularization

The *posterior regularization* (PR) framework introduced by Graça et al. (2008), and later refined in Ganchev et al. (2010), is notable early work which applied explicit performance constraints to latent variable model objective functions. Most of this work focused on models for only two local random variables: data x_d and hidden variables π_d, without any explicit labels y_d. Mindful of this, we can naturally express the PR objective in our notation, explaining data x explicitly via an objective function and incorporating labels y only later in the performance constraints.

The PR approach begins with the same overall goals of the expectation-maximization treatment of maximum likelihood inference: frame the problem as estimating an approximate posterior $q(\pi_d \mid \hat{v}_d)$ for each latent variable set π_d, such that this approximation is as close as possible in KL divergence to the real (perhaps intractable) posterior $p(\pi_d \mid x_d, y_d, \phi, \eta)$. Generally, we select the density q to be from a tractable parametric family with free parameters $\hat{v}_d$ restricted to some parameter space $\hat{v}_d \in \mathcal{V}$

which makes q a valid density. This leads to the objective

$$\min_{\phi, \{\hat{v}_d\}_{d=1}^D} \quad -\log p(\phi) - \sum_{d=1}^{D} \mathcal{L}(x_d, \hat{v}_d, \phi), \tag{5}$$

$$\mathcal{L}(x_d, \hat{v}_d, \phi) \triangleq \mathbb{E}_q\Big[\log p(x_d, \pi_d \mid \phi) - \log q(\pi_d \mid \hat{v}_d)\Big] \leq \log p(x_d|\phi). \tag{6}$$

Here, the function $\mathcal{L}$ is a strict *lower bound* on the data likelihood $\log p(x_d \mid \phi)$. The popular EM algorithm optimizes this objective via coordinate descent steps that alternately update variational parameters $\hat{v}_d$ and model parameters ϕ. The PR framework of Graça et al. (2008) adds additional constraints to the approximate posterior $q(\pi_d \mid \hat{v}_d)$ so that some additional loss function of interest, over both observed and latent variables, has bounded value under the distribution $q(\pi_d)$:

$$\text{Posterior Regularization (PR):} \quad \mathbb{E}_{q(\pi_d)}\Big[\text{loss}(y_d, \hat{y}(x_d, \pi_d, \eta))\Big] \leq L. \tag{7}$$

For our purposes, one possible loss function could be the negative log likelihood for the label y: $\text{loss}(y_d, \hat{y}(x_d, \pi_d, \eta)) = -\log p(y_d \mid \pi_d, \eta)$. It is informative to directly compare the PR constraint above with the PC objective of Eq. (2). Our approach directly constrains the expected loss under the *true* hidden-variable-from-data posterior $p(\pi_d|x_d)$:

$$\text{Prediction Constrained (PC):} \quad \mathbb{E}_{p(\pi_d|x_d)}\Big[\text{loss}(y_d, \hat{y}(x_d, \pi_d, \eta))\Big] \leq L. \tag{8}$$

In contrast, the PR approach in Eq. (7) constrains the expectation under the *approximate posterior* $q(\pi_d)$. This posterior does not have to stay close to *true* hidden-variable-from-data posterior $p(\pi_d|x_d)$. Indeed, when we write the PR objective in unconstrained form with Lagrange multiplier λ, and assume the loss is the negative label log-likelihood, we have:

$$\min_{\phi, \eta, \{\hat{v}_d\}_{d=1}^D} -\mathbb{E}_q\left[\sum_{d=1}^{D} \log p(x_d, \pi_d \mid \phi) + \lambda \log p(y_d \mid \pi_d, \eta) - \log q(\pi_d|\hat{v}_d)\right] - \log p(\phi, \eta) \tag{9}$$

Shown this way, we reach a surprising conclusion: the PR objective reduces to a lower bound on the symmetric joint likelihood with labels replicated λ times. Thus, it will inherit all the problems of label replication discussed above, as the optimal training update for $q(\pi_d)$ incorporates information from *both* data x_d and labels y_d. However, this does *not* train the model to find topics ϕ which lead to good estimates of the asymmetric predictive density of labels given data $p(y_d \mid x_d, \phi, \eta)$, which we show is critical for good predictive performance.

A.5 Advantages over maximum entropy discrimination and regularized Bayes

Another key thread of related work putting constraints on approximate posteriors is known as *maximum entropy discrimination* (MED), first published in Jaakkola et al. (1999b) with further details in followup work (Jaakkola et al., 1999a; Jebara, 2001). This approach was developed for training discriminative models without hidden variables, where the primary innovation was showing how to manage uncertainty about parameter estimation under max-margin-like objectives. In the context of LVMs, this MED work differs from standard EM optimization in two important and separable ways. First, it estimates a posterior for global parameters $q(\phi)$ instead of a simple point estimate. Second, it enforces a margin constraint on label prediction, rather than just maximizing log probability of labels. We note briefly that Jaakkola et al. (1999a) did consider a MED objective for *unsupervised* latent variable models (see their Eq. 48), where the constraint is directly on the expectation of the lower-bound of the log data likelihood. The choice to constrain the data likelihood is fundamentally different from constraining the labels-given-data loss, which was not done for LVMs by the original MED work yet is more aligned with our focus with high-quality predictions.

The key application MED to supervised LVMs has been Zhu et al. (2012)'s MED-LDA, an extension of the LDA topic model based on a MED-inspired training objective. Later work developed similar objectives for other LVMs under the broad name of *regularized Bayesian inference* (Zhu et al., 2014). To understand these objectives, we focus on Zhu et al. (2012)'s original unconstrained training objectives for MED-LDA for both regression (Problem 2, Eq. 8 on p. 2246) and classification

(Problem 3, Eq. 19 on p. 2252), which can be fit into our notation[1] as follows:

$$\min_{q(\phi,\eta),\{\hat{v}_d\}_{d=1}^D} \quad \mathrm{KL}(q(\phi,\eta)||p_0(\phi,\eta)) - \mathbb{E}_{q(\phi,\eta)}\Big[\sum_{d=1}^D \mathcal{L}(x_d,\hat{v}_d,\phi)\Big] \\ +C\sum_{d=1}^D \mathrm{loss}(y_d, \mathbb{E}_{q(\phi,\eta,\pi_d)}[\hat{y}_d(x_d,\pi_d,\eta)])$$

Here $C > 0$ is a scalar emphasizing how important the loss function is relative to the unsupervised problem, $p_0(\phi,\eta)$ is some prior distribution on global parameters, and $\mathcal{L}(x_d,\hat{v}_d,\phi)$ is the same lower bound as in Eq. (5). We can make this objective more comparable to our earlier objectives by performing point estimation of ϕ,η instead of posterior approximation, which is reasonable in moderate to large data regimes, as the posterior for the global parameters ϕ,η will concentrate. This choice allows us to focus on our core question of how to define an objective that balances data x and labels y, rather than the separate question of managing uncertainty during this training. Making this simplification by substituting point estimates for expectations, with the KL divergence regularization term reducing to the log prior $R(\phi,\eta) = -\log p_0(\phi,\eta)$, and the MED-LDA objective becomes:

$$\min_{\phi,\eta,\{\hat{v}_d\}_{d=1}^D} \quad R(\phi,\eta) - \sum_{d=1}^D \mathcal{L}(x_d,\hat{v}_d,\phi) + C\sum_{d=1}^D \mathrm{loss}(y_d, \mathbb{E}_{q(\pi_d)}[\hat{y}_d(x_d,\pi_d,\eta)]). \tag{10}$$

Both this objective and Graça et al. (2008)'s PR framework consider expectations over the approximate posterior $q(\pi_d)$, rather than our choice of the data-only posterior $p(\pi_d|x_d)$. However, the key difference between MED-LDA and the PR objectives is that the MED-LDA objective computes the loss of an expected prediction ($\mathrm{loss}(y_d, \mathbb{E}_q[\hat{y}_d])$), while the earlier PR objective in Eq. (7) penalizes the full expectation of the loss ($\mathbb{E}_{q(\pi_d)}[\mathrm{loss}(y_d,\hat{y}_d)]$). Earlier MED work (Jaakkola et al., 1999a) also suggests using an expectation of the loss, $\mathbb{E}_{q(\phi,\pi_d)}[\mathrm{loss}(y_d,\hat{y}_d(x_d,\pi_d,\eta))]$. Decision theory argues that the latter choice is preferable when possible, since it should lead to decisions that better minimize loss under uncertainty. We suspect that MED-LDA chooses the former only because it leads to more tractable algorithms for their chosen loss functions.

Motivated by this decision-theoretic view, we consider modifying the MED-LDA objective of Eq. (10) so that we take the full expectation of the loss. This swap can also be justified by assuming the loss function is *convex*, as are both the epsilon-insensitive loss and the hinge loss used by MED-LDA, so that Jensen's inequality may be used to bound the objective in Eq. (10) from above. The resulting training objective is:

$$\min_{\phi,\eta,\{\hat{v}_d\}_{d=1}^D} \quad R(\phi,\eta) - \sum_{d=1}^D \mathcal{L}(x_d,\hat{v}_d,\phi) + C\sum_{d=1}^D \mathbb{E}_{q(\pi_d)}\Big[\mathrm{loss}(y_d,\hat{y}_d(x_d,\pi_d,\eta))\Big]. \tag{11}$$

In this form, we see that we have recovered the symmetric maximum likelihood objective with label replication from Eq. (4), with y replicated C times. Thus, even this MED effort fails to properly handle the asymmetry issue we have raised, possibly leading to poor generalization performance.

[1] We note an irregularity between the classification and regression formulation of MED-LDA published by Zhu et al. (2012): while classification-MED-LDA included labels y only the loss term, the regression-MED-LDA included *two* terms in the objective that penalize reconstruction of y: one inside the likelihood bound term $\mathcal{L}$ as well as inside a separate epsilon-insensitive loss term. Here, we assume that only the loss term is used for simplicity.

B Dataset Descriptions

B.1 Toy Data: 3x3 Bars with Misspecified Labels

Our toy data analysis task is designed to illustrate why many existing methods fail to address the *asymmetry* and *cardinality imbalance* while our proposed PC-sLDA method succeeds.

We purposely construct this dataset so the sLDA topic model is somewhat *misspecified*: the best topics according to the unsupervised or supervised maximum likelihood objective perform little better than chance, but there is *some* word-cooccurance structure that can be found while still attaining perfect label predictions. The data exhibits cardinality imbalance: there are about 50 words per document in x_d but only one binary label. More importantly, it exhibits *asymmetric embedding*: for many proposed topics and weights, there is *huge* difference between π_d estimated when both x and y are observed (training mode), and when only x is observed (prediction mode).

This toy domain has 9 possible vocabulary terms, canonically arranged in a 3×3 square grid. We generate each document via a 3-step process. First, we generate data count vectors x_d, each of size 40 to 60 words, by drawing from exactly one or two of the $K = 4$ horizontal or vertical "bar" topics shown in the main paper (Fig. 1). None of these bars emit the top left corner word (vocab index 1 of 9), so at this stage only $x_{d,2:9}$ can be non-zero, where non-zero values of 8-22 are typical. Second, we generate the label y_d *independently* of $x_{d,2:9}$, by flipping a biased coin (20% probability of positive label). Finally, for those documents that are positive, we deterministically set $x_{d,1} = 1$. Thus, the top-left corner word is an unambiguous signal of the document's target label, but is perfectly uncorrelated with any other words. Furthermore, its typical count is much less than other words.

The final corpus includes 500 training documents, 500 validation documents, and 500 test documents. Each document has between 40 and 60 tokens. Example documents with positive labels $y_d = 1$ and negative labels $y_d = 0$ are shown in Fig. B.1.

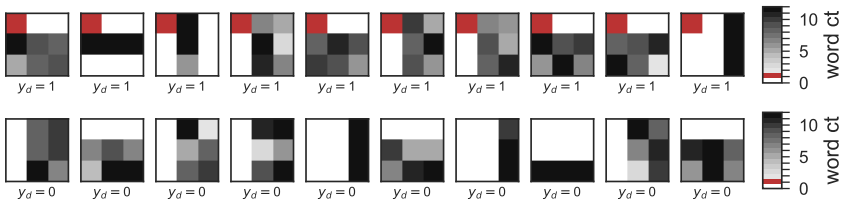

Fig. B.1: Example documents d for 3x3 bars task, shown with associated binary labels y_d (top row has $y_d = 1$, bottom row has $y_d = 0$). Note the special colormap chosen to highlight that the "signal" word (top left corner) appears with either a count of 1 or 0, while other words when present appear at much higher counts ≥ 10. Thus, generative objectives will often favor modeling these other words, while the signal word is the only reliable feature in the label prediction task.

Dataset access. Our curated labeled dataset and the code required to reproduce it can be obtained from our public codebase: `https://github.com/dtak/prediction-constrained-topic-models/datasets/toy_bars_3x3/`

B.2 Movie reviews

Raw text from movie reviews of four critics comes from scaledata v1.0 dataset released by Pang et al (Pang and Lee, 2005)[2]. Given plain text files of movie reviews, we tokenized and then stemmed using the Snowball stemmer from the nltk Python package, so that words with similar roots (e.g. film, films, filming) all become the same token. We removed all tokens in Mallet's list of common English stop words as well as any token included in the 1000 most common first names from the US census. We added this step after seeing too many common first names like Michael and Jennifer appear meaninglessly in many top-word lists for trained topics. We manually whitelisted "oscar" and "tony" due to their saliency to movie reviews sentiment. We then performed counts of all remaining tokens across the full raw corpus of 5006 documents, discarding any tokens that appear at least once

[2] `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

in more than 20% of all documents or less than 30 distinct documents. The final vocabulary list has 5375 terms.

Each of the 5006 original documents was then reduced to this vocabulary set. We discarded any documents that were too short (less than 20 tokens), leaving 5005 documents. Each document has a binary label, where 0 indicates it has a negative review (below 0.6 in the original datasets' 0-1 scale) and 1 indicates positive review (>= 0.6). This 0.6 threshold matches a threshold previously used in the raw data's 4-category scale to separate 0 and 1 star reviews from 2 and 3 (of 3) star reviews. Data pairs (x_d, y_d) were then split into training, validation, test. Both validation and test used 10 % of all documents, evenly balancing positive and negative labeled documents. The remaining documents were allocated to the training set.

Label statistics. The frequency of each binary label in the Movies training set is given in the table below:

attribute	fraction with attribute	count with attribute
more_than_2_out_of_4_stars	0.578	2315/4004

Word statistics. The Movies task training set's words-per-document statistics are in the table below:

	0%	1%	10%	50%	90%	99%	100%
unique tokens per doc	29	69	103	151	205	295	438
total tokens per doc	29	77	120	183	260	403	644

Dataset access. Our curated version of this dataset and the code required to reproduce it can be obtained from our public codebase: `https://github.com/dtak/prediction-constrained-topic-models/datasets/movie_reviews_pang_lee/`

B.3 Yelp reviews

We use raw text of online Yelp reviews from the Yelp dataset challenge (Yelp Dataset Challenge, 2016) to construct a multi-label binary dataset. This dataset includes text reviews about businesses. The businesses have associated meta data. We consider only businesses who have values for seven interesting binary attributes: "reservations accepted", "deliver offered", "alcohol served", "good for kids", "price range > 1" [3], "outdoor seating" and "wifi offered".

To construct the documents, we concatenate all reviews about a single business. Thus, each business is represented by a single document. We also prune the vocabulary, removing rare words that occur in fewer than 5 documents and removing very common words that occur in more than 50% of the documents. Finally, we sort the remaining words by tf-idf score and keep the top 10,000 scoring words as our final vocabulary.

The resulting corpus includes over 29,000 documents (23159 training, 2895 validation, and 2895 test) and a total of 43,236,060 observed words.

Label statistics. The frequency of each binary label in the Yelp training set is given in the table below:

attribute	fraction with attribute	count with attribute
reservations	0.418	9690/23159
delivery	0.206	4774/23159
alcohol	0.551	12762/23159
kid_friendly	0.835	19327/23159
expensive	0.610	14127/23159
outdoor_seating	0.416	9628/23159
wifi	0.428	9907/23159

Word statistics. The training set's words-per-document statistics are in the table below:

[3] Price range is given as an integer 1-4 where 1 is very cheap and 4 is very expensive. We turn this into a binary attribute by separating price range 1 from higher price ranges 2, 3 and 4.

	0%	1%	10%	50%	90%	99%	100%
unique tokens per doc	9	35	112	464	1612	3445	7461
total tokens per doc	9	41	147	772	4178	16794	213264

Dataset access. The terms of the Yelp dataset release prevent us from sharing our curated labeled dataset directly. However, we will happily share our preprocessing code to help interested parties recreate our curated dataset from the original 2016 data release. Please contact the first author via email: `mike@michaelchughes.com`.

B.4 Antidepressant Electronic Health Record (EHR) Dataset

We studied a broad cohort of hundreds of thousands patients drawn from a large academic medical center in New England and its affiliated outpatient network over a period of several years between 1997 and 2014. The cohort focused on individuals between age 18 and 80 who had at least one ICD9 diagnostic code for major depressive disorder (ICD9 codes 296.2x or 3x or 311). Our institutional review board approved the study protocol, waiving the requirement for informed consent.

From this broad cohort, we extracted all ICD-9 diagnostic codes, CPT procedure codes, and inpatient and outpatient medication prescriptions to represent patient history via a bag-of-codewords. We then identified a subset of patients who met a definition of *stable treatment* using a list of common anti-depressants marked as "primary" treatments for major depressive disorder by clinical collaborators. We labeled a treatment interval of a patient's record "stable" if all prescription events in the interval used the same subset of primary drugs, the interval lasted at least 90 days, and encounters occurred at least every 13 months.

Applying this criteria, we identified 29774/3721/3722 (training/validation/test) patients who met our stable treatment definition and also had sufficient history (a record containing at least two events before the first MDD prescription).

For each patient, we extracted a bag-of-codewords x_d of 5126 possible codewords (representing medical history before any stable treatment) and binary label vector y_d, marking which of 11 prevalent anti-depressants (if any) were used in known stable treatment.

Label statistics. The frequency of each binary label in the Antidepressant task training set is given in the table below:

attribute	**fraction with attribute**	**count with attribute**
nortriptyline	0.029	850/29774
amitriptyline	0.038	1138/29774
bupropion	0.139	4132/29774
fluoxetine	0.167	4972/29774
sertraline	0.155	4609/29774
paroxetine	0.080	2392/29774
venlafaxine	0.054	1606/29774
mirtazapine	0.032	961/29774
citalopram	0.252	7496/29774
escitalopram	0.050	1499/29774
duloxetine	0.029	853/29774

Word statistics. The Antidepressant task training set's words-per-document statistics are in the table below:

	0%	1%	10%	50%	90%	99%	100%
unique tokens per doc	1	2	8	61	194	375	829
total tokens per doc	2	2	17	195	968	2721	11317

Public release: Unfortunately, due to privacy concerns this dataset cannot be made public. For specific questions or concerns, please contact the first author via email: `mike@michaelchughes.com`.

C Details of Experimental Protocol

We outline several key decisions used in the experimental protocol. For full scripts necessary to reproduce our analyses, see our public code base online:

`https://github.com/dtak/prediction-constrained-topic-models/`

Duration. Methods were allowed to run for 5000 complete passes through the dataset, or up to 48 hours, whichever came first.

Initialization. For non-toy datasets, we consider two possible ways to initialize topics ϕ.

First, we construct a "from scratch" random initialization which draws all topics from low-variance random noise so no initial topic is too extreme yet symmetry breaking occurs. This initialization was used by all inference methods.

Second, we initialize to the learned topics produced by the Gibbs sampler for unsupervised LDA, taking the best scoring model snapshot on the validation set according to the score function outlined in the main paper. This was only used by our PC methods, but could easily be used by all other methods.

Batch sizes. We used different batch sizes for different data sets, as they were of different sizes and analysed on different computing architectures with different capabilities:

- Movie reviews: 1 batch (4004 docs / batch).
- Yelp: 20 batches (1157 docs / batch)
- Psychiatric EHR: 20 batches (1488 docs / batch).

Step sizes. PC-sLDA requires the choice of step size for the Adam gradient descent optimizer. We grid search among values between 0.001 and 0.1. Generally, larger values are preferred.

BP-sLDA also requires a step size for its mirror descent algorithm. We grid search among values between 0.001 and 0.1.

D Extended Results

D.1 Extended Results: 3x3 Bars

For the 3x3 bars task described in Fig. 1 of the main paper, we have several supplemental results.

Best parameter visualizations. First, we show extended visualizations of the final topic-word parameters and corresponding regression weights learned for each training method in Fig. D.1

By scanning Fig. D.1, we can immediately see that PC $\lambda \geq 10$ methods are the only ones which use a dedicated topic (the last topic in the plot's left-to-right order) to explain the top-left-corner signal word.

Differences are due to fundamental differences in the objectives, not lucky initialization. Next, in Fig. D.2 we examine the impact of various hand-designed initializations on the outcomes of different training methods to show that the differences between the training methods are *not* due to poor exploration of the optimization space, but are instead due to fundamental differences in the preferred parameters of each objective.

Fig. D.2 considers two distinct initializations which give better discriminative performance: top row is good discriminative but poor generative performance, bottom row is a balance of good discriminative *and* good generative performance (a configuration which would be good under our PC objective). We see that under both initializations, supposedly "supervised" methods like sLDA and MED-sLDA which do not account for the fundamental asymmetry of predicting labels from data wander away from these high-performing discriminative initializations. In contrast, our PC approach reaches the ideal lower left corner (good data likelihood and good label predictions) in both cases.

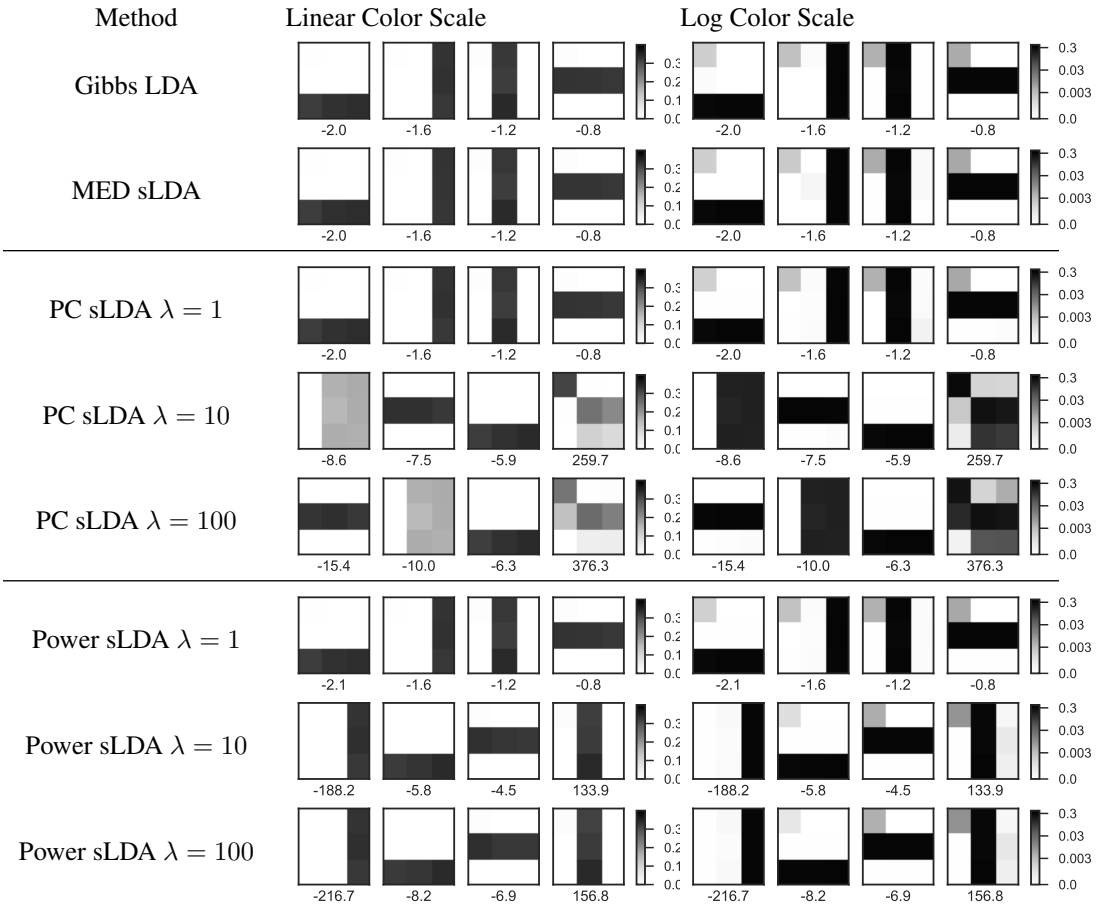

Fig. D.1: Extended results of parameter visualizations for the toy-bars task.

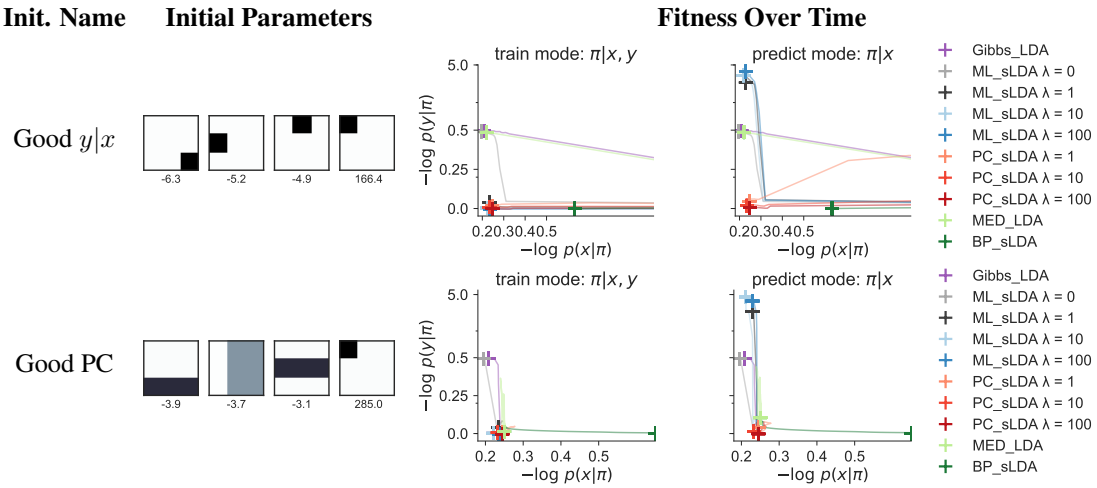

Fig. D.2: Extended toy bars task results: Evolution of topics from fixed initializations. Crosses mark the location of each method's training run which minimizes its training objective. *Top Row:* We consider initial topic-word parameters (left panel) designed to have good discriminative likelihood but terrible data likelihood. Right panel shows the trace of each method's location in the fitness landscape throughout training, using this "from good $y|x$" initialization and running until convergence. We see all methods evolve away from the initial configuration, with only PC methods with $\lambda \geq 10$ reaching the ideal lower corner of the fitness space. *Bottom Row:* We consider initial topic-word parameters (left panel) designed to have good scores under our PC objective. Right panel shows the trace of each method's location in the fitness landscape throughout training, using the "from good PC" initialization and running until convergence. We see all methods but high-λ PC-sLDA wander away from the initial configuration.

D.2 Extended Results: Movies

We next show extended results of our semisupervised experiments on the Movies dataset. This is like Fig. 2 from the main paper, but with additional K values across the full range $K = 10, 25, 50, 100$.

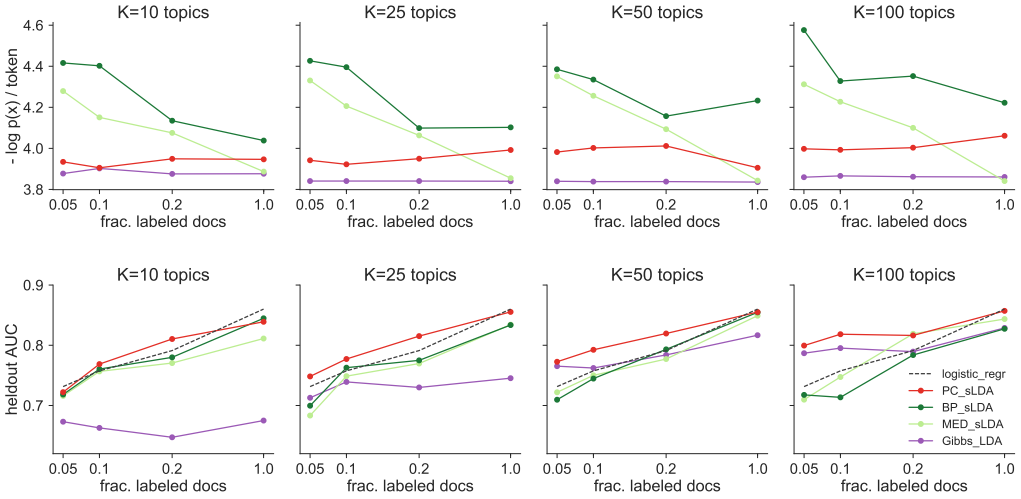

Fig. D.3: Extended results on Movie task.

D.3 Extended Results: Yelp

We next show extended results of our semisupervised experiments on the Yelp dataset. This is like Fig. 2 from the main paper, but with additional K values across the full range $K = 10, 25, 50, 100$.

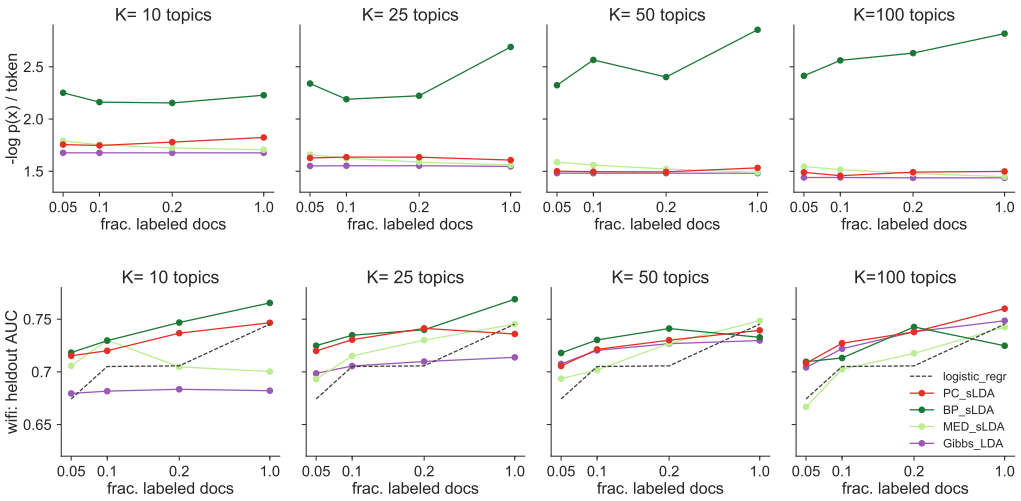

Fig. D.4: Extended results on Yelp task.

D.4 Extended Results: Antidepressant task

Here, we show (as claimed in the main paper) that we can use our code to train a BP-sLDA model (called "ourBPsLDA"), by setting the weight in front of our data likelihood to zero. The resulting code, when run on the big cohort of major depression disorder (MDD) patients, rapidly escapes from the Gibbs initialization and severely overfits, improving training loss at the expensive of heldout sets.

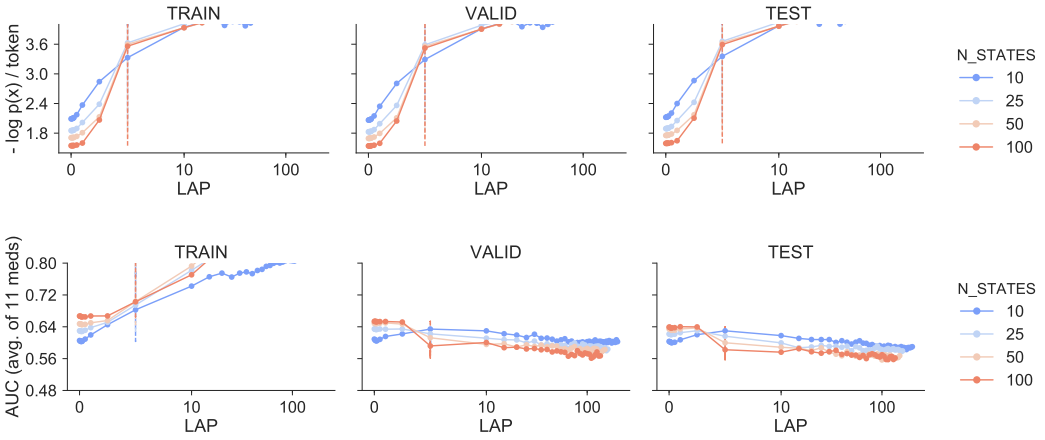

Fig. D.5: Extended results of our implementation of BP-sLDA (which can accomodate multiple binary labels) on the Antidepressant task.

E Browseable Visualizations of Learned Topics

We have created browseable HTML visualizations of the trained topic-word parameters ϕ and regression weights η for all datasets.

Simply point your favorite web browser to the links below to browse

- Movies $K = 25$
- Yelp $K = 25$
- Antidepressant $K = 25$

Once on these pages, you can click different links to explore different model sizes K and different label prediction coefficients within the selected task.

Two possible views of the top words for each topic are available:

- Folders marked "rerank_word=0" provide the classic view of a topic's top-word list. Each topic's words are sorted by $p(\text{word}|\text{topic})$.
- Folders marked "rerank_word=1" provide an alternative that identifies *anchor words*, that is, words whose presence in a document most strongly signals to use that topic. In these plots, each topic's words are sorted by $p(\text{topic}|\text{word})$.

All HTML files for these visualizations are available as a .zip file for download if you'd like to browse locally:

`http://michaelchughes.com/public_html/aistats_topic_viz_html.zip`

References

C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

J. Chen, J. He, Y. Shen, L. Xiao, X. He, J. Gao, X. Song, and L. Deng. End-to-end learning of LDA by mirror-descent back propagation over a deep architecture. In *Neural Information Processing Systems*, 2015.

K. Ganchev, J. Graça, J. Gillenwater, and B. Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, Aug. 2010.

J. Graça, K. Ganchev, and B. Taskar. Expectation maximization and posterior constraints. In *Neural Information Processing Systems*, 2008.

T. S. Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. Technical Report AITR-1668, Artificial Intelligence Laboratory at the Massachusetts Institute of Technology, 1999a. URL `http://people.csail.mit.edu/tommi/papers/maxent.ps`.

T. S. Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. In *Neural Information Processing Systems*, 1999b.

T. Jebara. *Discriminative, generative and imitative learning*. PhD thesis, Massachusetts Institute of Technology, 2001.

T. Jebara and A. Pentland. Maximum conditional likelihood via bound maximization and the CEM algorithm. In *Neural Information Processing Systems*, 1999.

J. D. McAuliffe and D. M. Blei. Supervised topic models. In *Neural Information Processing Systems*, pages 121–128, 2008.

B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, 2005.

Y. Ren, Y. Wang, and J. Zhu. Spectral learning for supervised topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

C. Wang, D. Blei, and F.-F. Li. Simultaneous image classification and annotation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

Y. Wang and J. Zhu. Spectral methods for supervised topic models. In *Advances in Neural Information Processing Systems*, pages 1511–1519, 2014.

Yelp Dataset Challenge. Yelp dataset challenge. `https://www.yelp.com/dataset_challenge`, 2016. Accessed: 2016-03.

C. Zhang and H. Kjellström. How to supervise topic models. In *ECCV Workshop on Graphical Models in Computer Vision*, 2014.

J. Zhu, A. Ahmed, and E. P. Xing. MedLDA: maximum margin supervised topic models. *The Journal of Machine Learning Research*, 13(1):2237–2278, 2012.

J. Zhu, N. Chen, and E. P. Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *Journal of Machine Learning Research*, 15(1):1799–1847, 2014.