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 Sudderth[2,3], and Finale Doshi-Velez[1]
[1]School of Engineering and Applied Sciences, Harvard University
[2]School of Information & Computer Sciences, Univ. of California, Irvine
[3]Department of Computer Science, Brown University [4]Massachusetts General Hospital & Harvard Medical School

Abstract

Supervisory signals can help topic models discover low-dimensional data representations which are useful for a specific prediction task. We propose a framework for training supervised latent Dirichlet allocation that balances two goals: faithful generative explanations of high-dimensional data and accurate prediction of associated class labels. Existing approaches fail to balance these goals by not properly handling a fundamental asymmetry: the intended application is always predicting labels from data, not data from labels. Our new *prediction-constrained* objective for training generative models coherently integrates supervisory signals even when only a small fraction of training examples are labeled. We demonstrate improved prediction quality compared to previous supervised topic models, achieving results competitive with high-dimensional logistic regression on text analysis and electronic health records tasks while simultaneously learning interpretable topics.

1 Introduction

Discrete count data are common: news articles can be represented as word counts, patient records as diagnosis counts, and images as visual descriptor counts. Topic models such as *latent Dirichlet allocation* (LDA, Blei et al. (2003)) are popular for finding cooccurance structure in datasets of count vectors, producing a small set of learned *topics* which help users understand the core themes of a corpus too large to comprehend manually. The low-dimensional feature space produced by LDA is often used as the input to some predictive task, where the user seeks to predict *labels* associated with each count vector. For example, Paul and Dredze (2014) use topics from Twitter to model flu trends, and Jiang et al. (2015) use topics from image captions to make travel recommendations. This approach balances two distinct goals: building a reasonable density model of observed data and making high-quality predictions of target labels. If we only cared about modeling the data well, we could simply apply standard topic models and ignore the labels. If we only cared about prediction performance, there are a host of well-studied supervised learning methods that we could apply directly to the labeled count vectors. However, using LDA-based dimensionality reduction as input to a predictor can simultaneously achieve two useful goals: good predictions of labels from count vectors *and* interpretable low-dimensional data representations.

Unfortunately, the two-stage pipeline of training LDA from count vectors alone and then training a supervised predictor from learned topic features often fails to produce accurate predictions. This is especially true when the count features are not carefully curated and contain structure irrelevant to the target task. For example, applying LDA to clinical records might find topics about common conditions like diabetes or heart disease, which may be irrelevant if the downstream task is predicting depression outcomes. To address this concern, many approaches have been developed to *supervise* topic models (McAuliffe and Blei, 2008; Lacoste-Julien et al., 2009; Zhu et al., 2012); the hope is that including task-specific labels during training will focus learned topics on the intended task, producing better predictions and more interpretable topics which ignore irrelevant parts of the data. However, a survey by Halpern et al. (2012) finds that existing supervised topic models offer little (if any) improvement in prediction quality over the baseline two-stage pipeline that trains topics without supervision, especially if the number of topics is larger than ∼10.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

In this work, we expose and correct key deficiencies in previous formulations of supervised topic models. We introduce a learning objective that directly encourages low-dimensional data representations to produce accurate predictions. Unlike earlier work, our training objective deliberately encodes the *asymmetry* of prediction tasks: web analysts want to predict WiFi availability tags from restaurant review text, not text from tags; clinicians want to predict medication outcomes given medical records, not records given outcomes. Approaches like *supervised LDA* (sLDA, McAuliffe and Blei (2008)) that optimize the *joint* likelihood of labels and words ignore this crucial asymmetry.

Our *prediction-constrained* (PC) latent variable models are tuned to maximize the marginal likelihood of the observed data, subject to the constraint that prediction accuracy (formalized as the conditional probability of labels given data) exceeds some target threshold. The PC objective addresses subtle but important limitations in nearly a decade of prior work on sLDA. We see clear empirical benefits from PC training: sometimes other objectives also work well, but ours *always* does.

In many applications, bag-of-words documents (text reviews, medical records, etc.) are plentiful and easy to collect. In contrast, reliable labels for these documents are expensive to obtain. Thus, it is beneficial for methods to be able to learn from datasets where only a small fraction of documents are labeled. For such semi-supervised learning, the difference between our PC objective and other approaches becomes more dramatic, and we see corresponding gains in performance.

2 Background: Topic Models

Standard LDA. The LDA topic model finds structure in a collection of D documents, or more generally, D examples of count vectors. Each document d is represented by a count vector x_d of V discrete words or features: $x_d \in \mathbb{Z}_+^V$. The LDA model generates these counts via a document-specific mixture of K topics:

$$\begin{aligned} \pi_d|\alpha &\sim \mathrm{Dir}(\pi_d \mid \alpha), \\ x_d|\pi_d, \phi &\sim \mathrm{Mult}(x_d \mid \textstyle\sum_{k=1}^K \pi_{dk}\phi_k, N_d). \end{aligned} \tag{1}$$

The random variable π_d is a document-topic probability vector, where π_{dk} is the probability of topic k in document d and $\sum_{k=1}^K \pi_{dk} = 1$. The vector ϕ_k is a topic-word probability vector, where ϕ_{kv} gives the probability of word v in topic k and $\sum_{v=1}^V \phi_{kv} = 1$. N_d is the (observed) size of document d: $N_d = \sum_v x_{dv}$. LDA assumes π_d and ϕ_k have symmetric Dirichlet priors, with hyperparameters $\alpha > 0$ and $\tau > 0$.

Topic-based Prediction of Binary Labels. Suppose document d also has a binary label $y_d \in \{0, 1\}$. Standard supervised topic models assume labels and word counts are conditionally independent given document-topic probabilities π_d:

$$y_d|\pi_d, \eta \sim \mathrm{Bern}(y_d \mid \sigma(\textstyle\sum_{k=1}^K \pi_{dk}\eta_k)), \tag{2}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ is the logit function, and $\eta \in \mathbb{R}^K$ is a vector of real-valued regression weights with a vague prior $\eta_k \sim \mathcal{N}(0, \sigma_\eta^2)$. Non-binary labels can be predicted via a generalized linear model (McAuliffe and Blei, 2008). In some experiments, we model vectors of binary labels $y_d \in \{0, 1\}^L$ with L conditionally independent logistic regressions.

The sLDA model of McAuliffe and Blei (2008) represents the count likelihood of Eq. (1) via N_d independent assignments $z_{dn} \sim \mathrm{Cat}(\pi_d)$ of word tokens to topics, and generates labels $y_d \sim \mathrm{Bern}(y_d \mid \sigma(\sum_{k=1}^K \bar{z}_{dk}\eta_k))$, where $\bar{z}_d = N_d^{-1} \sum_n z_{dn}$ and $E[\bar{z}_d] = \pi_d$. To enable more efficient inference algorithms, we analytically marginalize the topic assignments z_d away in Eq. (1,2).

There also exist "upstream" variants of supervised topic models (Lacoste-Julien et al., 2009; Mimno and McCallum, 2008) in which the document-topic probabilities π_d have a distribution that is conditioned on the label y_d. We focus on "downstream" topic models as in Eq. (2) because they are more easily learned from data in which not all documents have labels y_d.

3 Limitations of Existing Objectives

There are a host of training objectives and inference algorithms for supervised LDA, including (McAuliffe and Blei, 2008; Wang et al., 2009; Zhu et al., 2012; Chen et al., 2015). One core contribution of this work is to identify a fundamental shortcoming of all these objectives: they do not actually optimize models to perform well at the intended asymmetric prediction task of labels from words. This fundamental shortcoming arises due to model misspecification. If our count data truly came from a topic model, and those topics truly were the key to good predictions, then even a standard unsupervised topic model would do well. Trouble arises when we desire the dimensionality reduction provided by a topic model, for interpretability or efficiency, but the count data were not actually produced by the LDA generative process.

Limitations of Joint Bayesian or Maximum-Likelihood Training of the sLDA Model. The original formulation of supervised LDA (McAuliffe and Blei, 2008) and related work (Wang et al., 2009; Wang and Zhu, 2014; Ren et al., 2017) assumes a graphical model in which the target label y_d can be viewed as yet another output of document-topic probabilities π_d. When the number of counts in x_d is significantly larger than the cardinality of y_d, as is typical in practice, the likelihood associated with x_d will be much larger in magnitude than the likelihood associated with y_d. That is, the *correct* application of Bayesian inference for this joint likelihood model $p(x, y \mid \eta, \phi, \alpha)$ will produce learned topics that are *indistinguishable* from those estimated from the data likelihood $p(x|\phi, \alpha)$ alone.

Limitations of Label Replication. Many have observed that standard sLDA is insufficiently discriminative. The Power-sLDA approach of Zhang and Kjellström (2014) seeks to improve discriminative performance by "observing" $R > 1$ artificial copies of the y_d labels. Bayesian inference then focuses on the replicated likelihood $p(x, y, y, \ldots, y \mid \eta, \phi, \alpha)$. Unfortunately, the posterior $p(\pi_d \mid x, \phi)$ required to make predictions at test time may be very different from the posterior $p(\pi_d \mid x, y, \ldots, y, \phi, \eta)$ used at training; this can lead to low-dimensional representations π_d that fail to predict well at test time, when only x_d is observed. Fig. 1 demonstrates this issue on toy data: regardless of the replication level, Power-sLDA (blue)'s preferred model makes terrible predictions in a test setting using only x_d, even though the same model scores well at 'predictions' during training (when both x_d and y_d are observed). In the supplement, we expose the formal differences between label replication and our prediction-constrained objective in more detail.

Other Popular Objectives Reduce to Label Replication. Posterior regularization (PR, Ganchev et al. (2010); Graça et al. (2008)) enforces explicit performance constraints on the posterior. The MedLDA approach of Zhu et al. (2012, 2013, 2014) is derived from a maximum entropy discrimination framework and uses a hinge loss to penalize errors in the prediction of y_d. In the supplement, we show that after replacing the hinge loss with a logistic loss and approximate posteriors with point estimates, both MedLDA and PR can be written as forms of label replication. Thus, both can fail to learn topics that offer competitive predictions of labels from words at test time.

Limitations of Fully Discriminative Learning. Unlike all previous approaches, mirror-descent backpropagation sLDA (BP-sLDA, Chen et al. (2015)) focuses *entirely* on the prediction of y_d from x_d. Topics ϕ are trained to directly predict y_d from x_d via latent document-topic probabilities π_d, but no term in the objective ensures that topics ϕ accurately model x_d. Our objective can be seen as a generalization that balances the explanation of data x_d (which Chen et al. (2015) ignores) and prediction of targets y_d.

Partial Supervision. Many semi-supervised methods for general latent variable models optimize the joint likelihood of data x and labels y by either imputing (Nigam et al., 1998; Chang et al., 2007) or marginalizing (Kingma et al., 2014) missing labels. We have shown joint likelihood training for supervised topic models to have prediction quality similar to the unsupervised case even when *all* examples are labeled; accuracy will not improve when labels are rare. Other semi-supervised methods (Mann and McCallum, 2010) maximize conditional likelihood $\log p(y \mid x)$ only and thus do not learn useful generative models as we do.

Previous work on semi-supervised training of *topic models* seems to be scarce. Huh and Fienberg (2012) regularize distances in the document-topic space, but do not directly incorporate labels y in their objective for training topics ϕ. Xiang and Zhou (2014) train a naïve Bayes classifier using topics learned from a larger, unlabeled corpus, but make no use of labels when learning topics. Our semi-supervised approach should thus be of broad interest, because it ensures that labels y impact the learned topics ϕ even if they are only present for a small subset of documents.

Finally, some semi-supervised learning methods specialized for text data respect word order (Johnson and Zhang, 2015), unlike our bag-of-words approach. However, unlike our method, such approaches are not trained end-to-end and are not explicitly optimized to balance generative and discriminative performance.

4 Prediction-Constrained sLDA

We propose a novel, *prediction-constrained* (PC) objective that finds the best generative model for words x, while satisfying the *constraint* that topics ϕ must yield accurate predictions about labels y given x alone:

$$\begin{aligned} \min_{\phi,\eta} \quad & -\Big[\textstyle\sum_{d=1}^{D} \log p(x_d \mid \phi, \alpha)\Big] - \log p(\phi, \eta) \qquad (3) \\ \text{subject to} \quad & -\textstyle\sum_{d=1}^{D} \log p(y_d \mid x_d, \phi, \eta, \alpha) \leq \epsilon. \end{aligned}$$

The scalar ϵ is the highest aggregate loss we are willing to tolerate, and $p(\phi, \eta) = p(\phi)p(\eta)$ are independent priors used for regularization. There are many variations on this theme; for example, one could instead use a hinge loss as in Zhu et al. (2012). The structure of Eq. (3) matches the goals of a domain expert who wishes to explain as much of the data x as possible, while still making sufficiently accurate predictions of y.

Applying the Karush-Kuhn-Tucker conditions, we transform the inequality constrained objective in Eq. (3) to an equivalent unconstrained optimization problem:

$$\begin{aligned} \min_{\phi,\eta} - \sum_{d=1}^{D} \Big[\log p(x_d|\phi) + \lambda_\epsilon \log p(y_d|x_d, \phi, \eta) \Big] \qquad (4) \\ - \log p(\phi, \eta). \end{aligned}$$

For any prediction tolerance ϵ, there exists a scalar multiplier $\lambda_\epsilon > 0$ such that the optimum of Eq. (3) is a minimizer of Eq. (4). The relationship between λ_ϵ and ϵ is monotonic, but does not have an analytic form; we must search over the one-dimensional space of penalties λ_ϵ for an appropriate value.

While our PC objective is superficially similar to Power-sLDA (Zhang and Kjellström, 2014) and MedLDA (Zhu et al., 2012), it is distinct: the multiplier λ_ϵ rescales the label posterior $p(y_d \mid x_d)$, while label-replication only upweights the label likelihood $p(y_d \mid \pi_d)$. By "replicating" the *entire* y-from-x posterior, our PC objective achieves our goal of accurately predicting targets from words alone at test time. The constraint in Eq. (3) also theoretically justifies the use of a replication weight.

Computing $p(x_d \mid \phi)$ and $p(y_d \mid x_d, \phi, \eta)$ requires the marginalization of π_d over its simplex domain Δ^K:

$$p(x_d|\phi) = \int \mathrm{Mult}(x_d|\textstyle\sum_{k=1}^K \pi_{dk}\phi_k)\mathrm{Dir}(\pi_d|\alpha)d\pi_d, \quad (5)$$

$$p(y_d|x_d, \phi, \eta) = \int \mathrm{Bern}(y_d|\sigma(\pi_d^T \eta))p(\pi_d|x_d, \phi, \alpha)d\pi_d.$$

Because $p(y_d \mid x_d, \phi, \eta)$ integrates over $p(\pi_d \mid x_d, \phi)$, the posterior of π_d given *only* words x_d, our PC objective encodes the asymmetry of label prediction tasks.

Unfortunately, these integrals are intractable. To gain traction, we first contemplate an objective that *instantiates* π_d rather than marginalizing π_d away:

$$\begin{aligned} \min_{\pi,\phi,\eta} -\sum_{d=1}^{D} \Big[&\log p(\pi_d|\alpha) + \log p(x_d|\pi_d, \phi) \qquad (6) \\ &\lambda \log p(y_d|\pi_d, \eta)\Big] - \log p(\phi, \eta) \end{aligned}$$

As discussed above, solutions to this objective would lead to replicated joint training and its poor predictions of y_d given x_d alone. Since we wish to train under the same asymmetric conditions present at test time, where we have x_d but not y_d, we *fix* π_d to a deterministic embedding of the words x_d to the topic simplex. We choose this mapping to produce the *maximum a posteriori* (MAP) estimate of π_d given x_d: $\pi_d = \mathrm{argmax}_{\pi_d \in \Delta^K} \log p(\pi_d|x_d, \phi, \alpha)$. As we show in Sec. 5, this MAP estimate can be found deterministically via a tractable function: $\pi_d \leftarrow \mathrm{MAP}(x_d, \phi, \alpha)$.

Our chosen MAP embedding is a feasible approximation to the full posterior $p(\pi_d|x_d, \phi, \alpha)$ needed in Eq. (5), with approximation accuracy increasing as the number of observed words N_d grows. We can now write a *tractable* PC training objective for sLDA:

$$\begin{aligned} \mathcal{J}(\phi, \eta) = -\sum_{d=1}^{D} \Big[&\log p(\mathrm{MAP}(x_d, \phi, \alpha)|\alpha) \qquad (7) \\ &+ \log p(x_d|\mathrm{MAP}(x_d, \phi, \alpha), \phi) \\ &+ \lambda_\epsilon \log p(y_d|\mathrm{MAP}(x_d, \phi, \alpha), \eta)\Big] \\ &- \log p(\phi, \eta). \end{aligned}$$

While this objective is similar to BP-sLDA (Chen et al., 2015), the key difference is that the prediction constraint of Eq. (3) leads to a multiplier λ_ϵ that balances the generative and discriminative objectives. In contrast, Chen et al. (2015) consider only a fully unsupervised objective (labels y are ignored) and a fully supervised objective (the distribution of x is ignored). If documents are partially labeled, the objectives of Eq. (3) and (7) can be naturally generalized to only include prediction constraints for observed labels.

5 Inference & Learning for PC-sLDA

We first show how to evaluate the PC objective of Eq. (7) by describing an algorithm that computes the embedding $\mathrm{MAP}(x_d, \phi, \alpha)$. We then differentiate through the entire objective to allow gradient-based optimization of the topic-word probability vectors $\{\phi_k\}_{k=1}^K$ and regression coefficients $\{\eta_k\}_{k=1}^K$.

MAP via Exponentiated Gradient. Sontag and Roy (2011) define the document-topic MAP estimation problem for LDA as $\max_{\pi_d \in \Delta^K} \ell(\pi_d; x_d, \phi, \alpha)$, where

$$\ell(\pi_d; \ldots) = \log \mathrm{Mult}(x_d \mid \pi_d^T \phi) + \log \mathrm{Dir}(\pi_d \mid \alpha). \quad (8)$$

This problem is convex for $\alpha \geq 1$ and non-convex otherwise. For the convex case, they apply an *exponentiated gradient* algorithm (Kivinen and Warmuth, 1997) that iteratively rescales elements of the probability vector with exponentiated derivatives of the objective ℓ:

$$\begin{aligned} &\text{init: } \pi_d^0 \leftarrow \Big[\tfrac{1}{K} \cdots \tfrac{1}{K}\Big], \qquad (9) \\ &\text{repeat: } \pi_{dk}^t \leftarrow \frac{p_{dk}^t}{\sum_{j=1}^K p_{dj}^t}, \quad p_{dk}^t = \pi_{dk}^{t-1} \cdot e^{\nu \nabla \ell(\pi_{dk}^{t-1})}. \end{aligned}$$

With small enough step size $\nu > 0$, exponentiated gradient (EG) converges to the MAP solution. We define our embedding function $\pi_d \leftarrow \mathrm{MAP}(x_d, \phi, \alpha)$ to be the deterministic outcome of T EG iterations. In experiments, we use $T \approx 100$ and $\nu \approx 0.005$.

When $\alpha < 1$, the sparsity-promoting Dirichlet prior may lead to multimodal posteriors on the simplex $\pi_d \in \Delta^K$. But as noted by Taddy (2012), if we instead use a softmax (MacKay, 1997) representation of π_d (the natural parameters of the corresponding exponential family), the posterior is log-concave with a single mode. Elegantly, the softmax-basis MAP for a particular $\alpha < 1$ equals the simplex MAP estimate under a modified Dirichlet prior, $p(\pi_d \mid x_d, \phi, \alpha + 1)$. Using this "add one" trick, exponentiated gradient gives optimal *natural parameter* MAP estimates even when $\alpha < 1$.

Parameter Learning via SGD. To optimize the objective in Eq. (7), we realize first that the iterative

MAP embedding in Eq. (9) is *differentiable* with respect to the parameters ϕ and η. This means the *entire* objective $\mathcal{J}$ is differentiable and modern gradient descent methods may be applied to learn ϕ, η from data, using standard transformations of constrained parameters ϕ from the simplex to the reals. Once the loss function is specified via unconstrained parameters, we perform automatic differentiation to compute gradients and then optimize via the Adam algorithm (Kingma and Ba, 2014). For scalability, we can perform stochastic updates from minibatches of data. We have developed Python implementations using both Autograd (Maclaurin et al., 2015) and Tensorflow (Abadi et al., 2015) which are available online.[1]

Previously, Chen et al. (2015) optimized a purely discriminative objective via mirror descent directly on the constrained parameters ϕ, using a C# implementation with manually-derived gradient computations. In contrast, our approach allows many useful extensions (such as multi-label binary classification) without need to derive and implement gradient calculations by hand.

Hyperparameter selection. The key hyperparameter for our PC-sLDA algorithm is the multiplier λ_ϵ. For topic models of count data, λ_ϵ usually should be on the order of the number of tokens in the average document, though it may need to be larger if tension exists between the unsupervised and supervised terms of the objective. As in our experiments, we suggest trying a logarithmically spaced range of values $\lambda_\epsilon \in \{10, 100, 1000, \ldots\}$. From this grid search, we select the value that minimizes a score defined later in Eq. (10) which assesses the model's combined discriminative and generative performance. The cost of multiple runs can be mitigated by using the final parameters at one λ_ϵ value as the initial parameters for the next run, although this may not escape to new preferred basins of attraction in the overall non-convex objective.

6 Experimental Results

We now assess how well our proposed PC training of sLDA (PC-sLDA) achieves its *simultaneous* goals of accurate prediction of labels y given x while maintaining faithful explanations of words x. Full descriptions of all datasets and procedures are in the supplement.

Tasks. We consider one toy and three real-world bag-of-words prediction tasks. For each non-toy dataset, we partition documents into three sets (training/validation/test). We tune hyperparameters on the validation set and report results on the test set.

- **Toy 3×3 Bars task.** To study trade-offs between models of $p(x)$ and $p(y|x)$, we built a toy dataset that is deliberately *misspecified*: neither the unsupervised LDA maximum likelihood solution nor the supervised sLDA maximum likelihood solution performs better than chance at label prediction. We look at 500 training documents, each with $V = 9$ possible vocabulary words that can be arranged in a 3-by-3 grid to indicate bar-like co-occurrence structure, as illustrated in Fig. 1. Each binary label y_d is unrelated to the observed x_d vector except for a rare signal word (top-left corner).
- **Movie task.** Each of the 4004/500/501 documents is a published movie review by a professional critic (Pang and Lee, 2005), with $V = 5338$ terms. Each review has one binary label, where $y_d = 1$ means the critic gave the film more than 2 of 4 stars.
- **Yelp task.** Each of the 23159/2895/2895 documents (Yelp Dataset Challenge, 2016) aggregates all text reviews for a single restaurant, using $V = 10,000$ vocabulary terms. Each document also has 7 possible binary labels y_d: takes-reservations, delivery, alcohol, good-for-kids, expensive, outdoor-patio, and wifi.
- **Antidepressant task.** Finally, we predict which subset of 11 common antidepressants would successfully treat an individual's major depressive disorder given a count vector x_d of the patient's electronic health record (EHR) code history. These are real de-identified data from tertiary care hospitals, split into 29774/3721/3722 documents (one per patient) with $V = 5126$ codewords which represent past diagnoses (ICD-9), procedures (CPT), and medications.

Baselines. Our discriminative baselines include logistic regression, the fully supervised BP-sLDA algorithm of Chen et al. (2015), the unsupervised Gibbs sampler for LDA (Griffiths and Steyvers, 2004) from the Mallet toolbox (McCallum, 2002), and the supervised MED-sLDA Gibbs sampler (Zhu et al., 2013) which is reported to improve on an earlier variational method (Zhu et al., 2012). To be fair to all methods, we tune relevant hyperparameters (L_2 regularization strength for regression, MED-sLDA regularization weight, step sizes, etc.) on validation data. For our toy example, we also compare to a coordinate ascent algorithm for the maximum-likelihood sLDA objective in Eq. (6) (Power-sLDA), across different values of the label replication factor $\lambda \geq 0$. Power-sLDA $\lambda = 0$ is equivalent to unsupervised LDA; Power-sLDA $\lambda = 1$ is the standard sLDA of McAuliffe and Blei (2008).

All baselines support documents with one binary label $y_d \in \{0, 1\}$. Third-party MED-sLDA and BP-sLDA code does not support multiple binary labels per document, but our PC-sLDA does. In these cases, we either train MED-sLDA on only one label (e.g., only wifi for the Yelp task) or omit it.

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

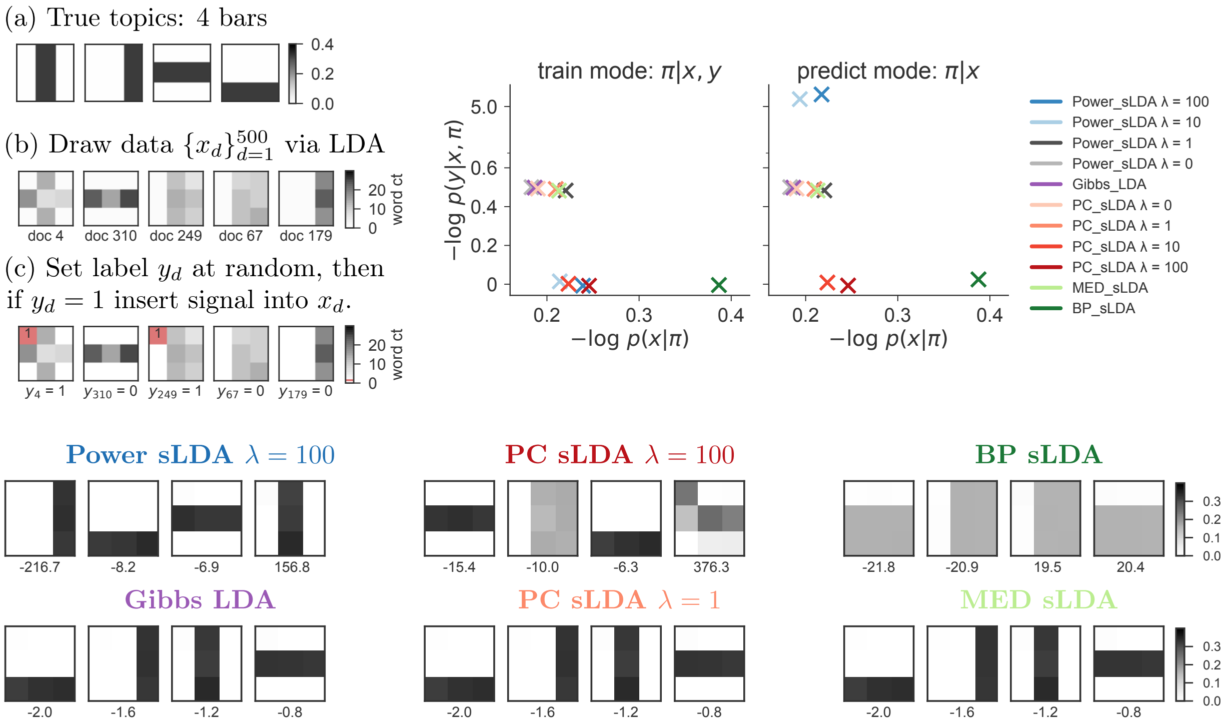


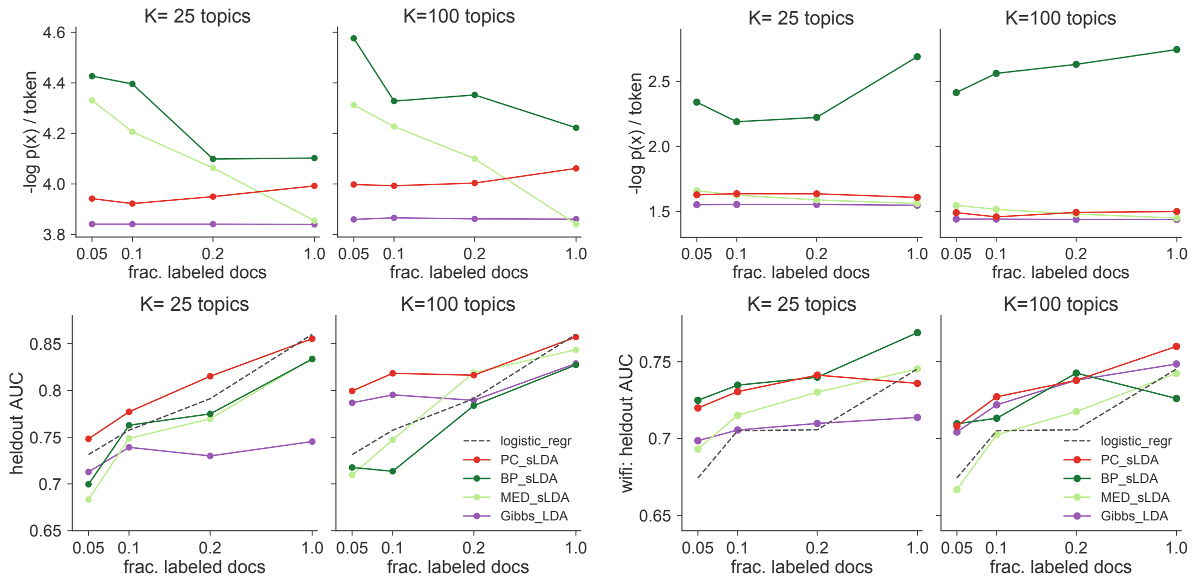
Figure 1: 3×3 bars task: The advantages of PC training under misspecification. Given only $K = 4$ topics, the goal is to simultaneously model the bar-like topic structure (as in Griffiths and Steyvers (2004)) of observed counts while making accurate binary label predictions using learned topic features. *Top Left:* Illustration of the true generative process for 5 example documents. Each document d has a binary label y_d and a count vector x_d over 9 possible vocabulary symbols arranged in a 3×3 square grid. To generate document d, we first draw x_d as a mixture of 4 true "bar" topics, as in LDA. Next, we draw $y_d \sim \text{Bern}(0.2)$, so it is *independent* of x_d and thus any sLDA model is misspecified. Finally, if $y_d = 1$ we set the top-left word $x_{d0} = 1$, otherwise $x_{d0} = 0$. Thus, there is a clear signal to predict y_d well given x_d but it relies on *none* of the bar topics. *Top Right:* Each method's best solution (as ranked by its training objective) is located on a 2-dimensional fitness landscape. The x-axis is negative log likelihood of data x averaged per token (lower is better). The y-axis is the negative log likelihood of labels y averaged per document (lower is better). These metrics are computed on the *training* set. We show these scores under two possible modes for estimating the document-topic vector π_d. *Train mode* finds the supervised MAP estimate $\max_{\pi_d} \log p(\pi_d | x_d, y_d, \phi, \eta, \alpha)$. *Predict mode* finds the unsupervised MAP estimate $\max_{\pi_d} \log p(\pi_d | x_d, \phi, \alpha)$. This distinction highlights the key difference between PC-sLDA with $\lambda > 1$, which deliberately trains topics to be good at labels-from-data prediction, and label replication (Power-sLDA with $\lambda > 1$), which trains models that do well in training mode but fail in a predictive setting (even on the same training data). *Bottom Rows:* Learned topic-word parameters for each method, labeled with regression coefficient η_k for each topic.

Partial Supervision. On Movie and Yelp tasks, we artificially include only a small fraction (0.05, 0.10, or 0.20) of available training labels, chosen at random. Fully supervised methods (e.g. BP-sLDA, MED-sLDA) are *only given* documents (x_d, y_d) from this subset, because third-party code does not allow using unlabeled data at training. Our PC-sLDA as well as Gibbs-LDA uses the entire partially-labeled training set.

Protocol. All topic models are run from multiple random initializations of ϕ, η (for fairness, all methods use same predefined initializations of these parameters). We record point estimates of topic-word parameters ϕ and regression weights η at defined intervals throughout training. For all methods, at each parameter snapshot ϕ, η we evaluate *discriminative* prediction quality via area-under-the-ROC-curve (AUC) using the predicted probability $\Pr(y_d = 1 | x_d) = \sigma(\eta^T \text{MAP}(x_d, \phi, \alpha))$. We evaluate *generative* model quality via a variational evidence lower bound on heldout per-token log likelihood: $(\sum_d N_d)^{-1} \sum_{d=1}^{D} \log p(x_d | \phi, \alpha)$. For all methods, we select the best snapshot on the validation set (early stopping) by minimizing the score:

$$-10 * \text{AUC}(y, x, \phi, \eta) - \text{PerTokELBO}(x | \phi, \alpha). \quad (10)$$

From-Gibbs Initializations. Our stochastic gradient descent algorithm for PC-sLDA is vulnerable to poor exploration of the non-convex space. To remedy this, we augment the randomly-initialized runs of PC-sLDA with separate initializations which start at the best parameter snapshot (ϕ, η) produced by unsupervised Gibbs-LDA. We then select the best PC-sLDA result among both from-random and from-Gibbs runs. This lets us assess the value of our new training objective without confounding due to poor initialization.



(a) Movie reviews

(b) Yelp reviews

Figure 2: Movie and Yelp tasks: Performance metrics vs. fraction of labeled training documents used for 25 and 100 topics. An extended version is in the supplement. *Top row:* Heldout generative performance (negative likelihood, lower is better). *Bottom row:* Heldout discriminative performance (AUC, higher is better). Note that improvements over supervised learning algorithms, including logistic regression, are particularly large when the fraction of labeled documents is small.

Hyperparameters. For non-toy tasks, we show performance across several model sizes $K \in \{10, 25, 50, 100\}$ (full results are in the supplement). We set the topic-word prior concentration $\tau = 0.01$ and grid search the document-topic prior concentration $\alpha \in \{0.1, 0.01\}$ and the regression weight variance $\sigma_\eta^2 \in \{5, 500\}$.

Results. Across all tasks, our major findings are:

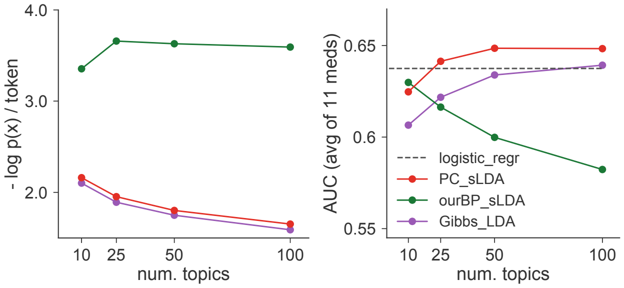
PC-sLDA has high-quality label prediction. When datasets are fully labeled, we sensibly find that purely discriminative methods like logistic regression (LR) or BP-sLDA often achieve the highest AUC values. But our PC-sLDA is consistently competitive, matching LR on the Movie task in Fig. 2, beating it slightly on the large-scale Antidepressant task in Fig. 3.

PC-sLDA is the only method robust to misspecification. In the toy bars task in Fig. 1, we see that our PC-sLDA with $\lambda \geq 10$ is the only method to find a topic with high probability on the signal word (top left corner), which is key to good discrimination. Most other methods, such as sLDA or MED-sLDA, are indistinguishable from the unsupervised LDA solution. Label replication (Power-sLDA $\lambda > 1$) suffers the most under misspecification, yielding solutions with terrible generalization performance. Purely discriminative BP-sLDA discriminates well but learns very poor generative models with no useful bar structure.

PC-sLDA predictions remain good when few documents have labels. For the Movie task in Fig. 2(a), PC-sLDA dominates the AUC metric for small fractions of labels (0.05, 0.1), beating even LR when K=100. In this regime, unsupervised Gibbs-LDA with $K = 100$ topics has better AUC than BP-sLDA and MED-sLDA, demonstrating the value of unlabeled data for prediction. On Yelp, PC-sLDA predictions at small fractions are better than all but BP-sLDA.

PC-sLDA recovers better heldout data likelihoods than BP-sLDA. Both Fig. 2 (top row) and Fig. 3 show trends in heldout data negative log likelihood (lower is better). As expected, unsupervised Gibbs-LDA consistently achieves the best scores, because explaining data is its sole objective. MED-sLDA also does reasonably, in some cases better than PC-sLDA, but usually in these cases MED-sLDA has worse AUC than PC-sLDA. BP-sLDA is consistently poor, having per-token likelihoods about 0.1-1.0 nats higher than others on full training sets. These results show that the solely discriminative approach of BP-sLDA cannot explain the data well. In contrast, our PC-sLDA can capture essential data properties while still predicting labels accurately.

PC-sLDA's learned topic-word probabilities ϕ are interpretable for prediction. A key point of our work is that our PC training estimates topic-word parameters ϕ to focus more on the label prediction than unsupervised training would. On the Antidepressant task, Fig. 3 shows that PC-sLDA initialized from Gibbs



Gibbs-LDA η_k = -1.6		PC-sLDA η_k =+0.3	
0.083	office visit >=15min	0.137	office visit >=15min
0.045	office visit >=05min	0.037	office visit >=05min
0.026	routine physical	0.030	office visit >=25min
0.025	office visit >=25min	0.029	routine physical exam
0.015	complete blood ct. test	0.021	complete blood ct. test
0.015	office visit >=10min	0.021	routine blood collection
0.013	flu vaccine	0.015	other malaise & fatigue
--		---	
1.000	meningococcal vaccine	0.987	test for strep throat
1.000	poliovirus vaccine	0.982	hidradenitis skin cond
0.999	routine physical age5-11	0.963	preventive counsel >=45min
0.999	typhoid vaccine	0.963	need for typhoid vaccine

Figure 3: Antidepressant prediction task. *Left:* Heldout negative likelihood (generative performance, lower is better). *Center:* Heldout AUC (discriminative, higher is better). We use our own implementation of BP-sLDA for this multiple binary label prediction task. Both PC-sLDA and BP-sLDA numbers here are the results of runs initialized from Gibbs. While BP-sLDA exhibits severe overfitting (see supplement), our PC-sLDA improves on the baseline Gibbs predictions reliably. *Right:* Comparison of topic #11 of $K = 25$ for both Gibbs-LDA and our PC-sLDA when initialized from Gibbs. We show the regression coefficient η_k for this topic when predicting patient success with drug citalopram. The top row is ranked by $p(\text{word}|\text{topic})$. The bottom row is ranked by $p(\text{topic}|\text{word})$, indicating potential *anchor words.* The original Gibbs topic is mostly about routine preventative care and vaccination. PC-sLDA training evolves the topic to emphasize longer duration encounters focused on counseling or behavior change, mixed together with a few infection words.

indeed causes an original Gibbs topic to significantly evolve its regression weight η_k and associated top words. The original Gibbs topic is mostly about routine outpatient preventative care and vaccination. The evolved PC-sLDA topic prefers long-duration primary care encounters focused on behavior change ("counseling"). With clinical collaborators, we hypothesize that this more focused topic leads to a positive η_k value because the drug citalopram is often a treatment of choice for such patients (i.e., uncomplicated MDD diagnosed and treated in primary care). The supplement contains browseable HTML visualizations of trained topic-word parameters for all datasets.

7 Discussion

Despite nearly a decade of work on supervised topic models, to our knowledge our prediction-constrained formulation is the only one that coherently manages the trade-off between modeling words and predicting labels. We demonstrate consistent advantages over baselines based on label replication, as these approaches fail to handle the asymmetry of the label prediction task. Discussing their original supervised LDA, Blei and McAuliffe (2010) say that a semi-supervised extension would be "straightforward," but caution that "care must be taken that the response [label] data exert sufficient influence on the fit." We have provided strong theoretical and empirical evidence that standard training of sLDA does *not* lead to effective semi-supervised learning nor good predictions, while PC-sLDA does.

Training. Prediction-constrained training of sLDA deviates from classical Bayesian methods due to the external requirement of good prediction performance. While this deviation may be unsettling, Liu et al. (2009) and Molitor et al. (2009) describe situations in which it is sensible to incorporate asymmetric inference strategies (cut distributions), which Plummer (2015) proves result in principled probability distributions that are not the posterior of any graphical model. The core question is what constraints should be introduced to solve the problem at hand. Ganchev et al. (2010), Graça et al. (2008), and Zhu et al. (2012, 2014) introduce distinct constraints that strengthen the link between document-topic vectors π_d and labels y_d. But as we showed in Sec. 3, such constraints do not necessarily improve label predictions from x_d alone.

Algorithm extensions. Deviating from the Bayesian framework means that many traditional inference tools are no longer applicable. We found that carefully-tuned gradient-based optimization could usually evolve to a good balance of generative and discriminative performance, especially given a good initialization from an LDA Gibbs sampler. Alternatively, continuation or homotopy methods (Corduneanu and Jaakkola, 2002) find a spectrum of models by smoothly varying λ_ϵ; in our case this means starting from an unsupervised model ($\lambda_\epsilon = 0$) and gradually increasing λ_ϵ while re-optimizing ϕ, η. From our preliminary experiments, it appears that the non-convex landscape across $\{\phi, \eta\}$ and λ_ϵ has sharp barriers: small changes in λ_ϵ can cause large changes in the optimal parameters $\{\phi, \eta\}$ which may not be reachable via gradual warm restarts. The development of improved inference algorithms is an exciting research direction.

Model extensions. In this paper, we have focused our algorithm design and experiments on the semi-supervised training of topic models. But by design, our prediction-constrained framework is directly applicable to an enormous range of latent variable models. We expect it to prove useful in many application domains.

Acknowledgements

This research supported in part by NSF CAREER Award No. IIS-1349774, and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA contract 2016-16041100004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government.

MCH is supported by Oracle Labs. We thank Victor Castro for preparing deidentified EHR data. We thank Harvard FAS Research Computing, Partners Research Computing, and Brown Computer Science for computational resources.

References

M. Abadi, A. Agarwal, P. Barham, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

D. M. Blei and J. D. McAuliffe. Supervised topic models. *arXiv preprint 1003.0783v1*, 2010.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003.

M.-W. Chang, L. Ratinov, and D. Roth. Guiding semi-supervision with constraint-driven learning. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, 2007.

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.

A. Corduneanu and T. Jaakkola. Continuation methods for mixing heterogeneous sources. In *Uncertainty in Artificial Intelligence*, 2002.

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. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 2004.

Y. Halpern, S. Horng, L. A. Nathanson, N. I. Shapiro, and D. Sontag. A comparison of dimensionality reduction techniques for unstructured clinical text. In *ICML workshop on clinical data analysis*, 2012.

S. Huh and S. E. Fienberg. Discriminative topic modeling based on manifold learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):20, 2012.

S. Jiang, X. Qian, J. Shen, and T. Mei. Travel recommendation via author topic model based collaborative filtering. In *International Conference on Multimedia Modeling*, pages 392–402. Springer, 2015.

R. Johnson and T. Zhang. Semi-supervised convolutional neural networks for text categorization via region embedding. In *Neural Information Processing Systems*, 2015.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint 1412.6980*, 2014.

D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Neural Information Processing Systems*, 2014.

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, 2009.

F. Liu, M. Bayarri, J. Berger, et al. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150, 2009.

D. J. C. MacKay. Ensemble learning for hidden Markov models. Technical report, Department of Physics, University of Cambridge, 1997.

D. Maclaurin, D. Duvenaud, M. Johnson, and R. Adams. Autograd: Reverse-mode differentiation of native python. `http://github.com/HIPS/autograd`, 2015.

G. S. Mann and A. McCallum. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, 11(Feb): 955–984, 2010.

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

A. K. McCallum. MALLET: Machine learning for language toolkit. `mallet.cs.umass.edu`, 2002.

D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, 2008.

N.-T. Molitor, N. Best, C. Jackson, and S. Richardson. Using Bayesian graphical models to model biases in observational studies and to combine multiple sources of data: application to low birth weight and water disinfection by-products. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3):615–637, 2009.

K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Learning to classify text from labeled and unla- beled documents. In *AAAI Conference on Artificial Intelligence*, 1998.

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.

M. J. Paul and M. Dredze. Discovering health topics in social media using topic models. *PLoS ONE*, 9(8), 2014.

M. Plummer. Cuts in Bayesian graphical models. *Statistics and Computing*, 25(1):37–43, 2015.

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

D. Sontag and D. Roy. Complexity of inference in latent dirichlet allocation. In *Neural Information Processing Systems*, 2011.

M. Taddy. On estimation and selection for topic models. In *Artificial Intelligence and Statistics*, 2012.

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*, 2014.

B. Xiang and L. Zhou. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, 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, H. Perkins, and B. Zhang. Gibbs max-margin topic models with fast sampling algorithms. In *International Conference on Machine Learning*, 2013.

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.