# Anchors Regularized: Adding Robustness and Extensibility to Scalable Topic-Modeling Algorithms

**Thang Nguyen**
iSchool and UMIACS,
University of Maryland
and National Library of Medicine,
National Institutes for Health
daithang@umiacs.umd.edu

**Yuening Hu**
Computer Science
University of Maryland
ynhu@cs.umd.edu

**Jordan Boyd-Graber**
iSchool and UMIACS
University of Maryland
jbg@umiacs.umd.edu

## Abstract

Spectral methods offer scalable alternatives to Markov chain Monte Carlo and expectation maximization. However, these new methods lack the rich priors associated with probabilistic models. We examine Arora et al.'s anchor words algorithm for topic modeling and develop new, regularized algorithms that not only mathematically resemble Gaussian and Dirichlet priors but also improve the interpretability of topic models. Our new regularization approaches make these efficient algorithms more flexible; we also show that these methods can be combined with informed priors.

## 1 Introduction

Topic models are of practical and theoretical interest. Practically, they have been used to understand political perspective (Paul and Girju, 2010), improve machine translation (Eidelman et al., 2012), reveal literary trends (Jockers, 2013), and understand scientific discourse (Hall et al., 2008). Theoretically, their latent variable formulation has served as a foundation for more robust models of other linguistic phenomena (Brody and Lapata, 2009).

Modern topic models are formulated as a latent variable model. Like hidden Markov models (Rabiner, 1989, HMM), each token comes from one of $K$ unknown distributions. Unlike a HMM, topic models assume that each document is an *admixture* of these hidden components called topics. Posterior inference discovers the hidden variables that best explain a dataset. Typical solutions use MCMC (Griffiths and Steyvers, 2004) or variational EM (Blei et al., 2003), which can be viewed as local optimization: searching for the latent variables that maximize the data likelihood.

An exciting vein of new research provides provable polynomial-time alternatives. These approaches provide solutions to hidden Markov models (Anandkumar et al., 2012), mixture models (Kannan et al., 2005), and latent variable grammars (Cohen et al., 2013). The key insight is not to directly optimize observation likelihood but to instead discover latent variables that can reconstruct statistics of the assumed generative model. Unlike search-based methods, which can be caught in local minima, these techniques are often guaranteed to find global optima.

These general techniques can be improved by making reasonable assumptions about the models. For example, Arora et al. (2012b)'s approach for inference in topic models assume that each topic has a unique "anchor" word (thus, we call this approach **anchor**). This approach is fast and effective; because it only uses word co-occurrence information, it can scale to much larger datasets than MCMC or EM alternatives. We review the **anchor** method in Section 2.

Despite their advantages, these techniques are not a panacea. They do not accommodate the rich priors that modelers have come to expect. Priors can improve performance (Wallach et al., 2009), provide domain adaptation (Daumé III, 2007; Finkel and Manning, 2009), and guide models to reflect users' needs (Hu et al., 2013). In Section 3, we regularize the **anchor** method to trade-off the reconstruction fidelity with the penalty terms that mimic Gaussian and Dirichlet priors.

Another shortcoming is that these models have not been scrutinized using standard NLP evaluations. Because these approaches emerged from the theory community, **anchor**'s evaluations, when present, typically use training reconstruction. In Section 4, we show that our regularized models can generalize to previously unseen data—as measured by held-out likelihood (Blei et al., 2003)—and are more interpretable (Chang et al., 2009; Newman et al., 2010). We also show that our extension to the **anchor** method enables new applications: for

| | |
|---|---|
| $K$ | number of topics |
| $V$ | vocabulary size |
| $M$ | document frequency: minimum documents an anchor word candidate must appear in |
| $\boldsymbol{Q}$ | word co-occurrence matrix $Q_{i,j} = p(w_1 = i, w_2 = j)$ |
| $\bar{\boldsymbol{Q}}$ | conditional distribution of $Q$ $\bar{Q}_{i,j} = p(w_1 = j \mid w_2 = i)$ |
| $\bar{\boldsymbol{Q}}_{i,\cdot}$ | row $i$ of $\bar{Q}$ |
| $\boldsymbol{A}$ | topic matrix, of size $V \times K$ $A_{j,k} = p(w = j \mid z = k)$ |
| $\boldsymbol{C}$ | anchor coefficient of size $K \times V$ $C_{j,k} = p(z = k \mid w = j)$ |
| $\mathcal{S}$ | set of anchor word indexes $\{s_1, \ldots s_K\}$ |
| $\lambda$ | regularization weight |

Table 1: Notation used. Matrices are in bold ($\boldsymbol{Q}, \boldsymbol{C}$), sets are in script $\mathcal{S}$

example, using an informed priors to discover concepts of interest.

Having shown that regularization does improve performance, in Section 5 we explore why. We discuss the trade-off of training data reconstruction with sparsity and why regularized topics are more interpretable.

## 2 Anchor Words: Scalable Topic Models

In this section, we briefly review the **anchor** method and place it in the context of topic model inference. Once we have established the **anchor** objective function, in the next section we regularize the objective function.

**Rethinking Data: Word Co-occurrence** Inference in topic models can be viewed as a black box: given a set of documents, discover the topics that best explain the data. The difference between **anchor** and conventional inference is that while conventional methods take a collection of documents as input, **anchor** takes *word co-occurrence* statistics. Given a vocabulary of size $V$, we represent this joint distribution as $\boldsymbol{Q}_{i,j} = p(w_1 = i, w_2 = j)$, each cell represents the probability of words appearing together in a document.

Like other topic modeling algorithms, the output of the **anchor** method is the topic word distributions $\boldsymbol{A}$ with size $V * K$, where $K$ is the total number of topics desired, a parameter of the algorithm. The $k^{th}$ column of $A$ will be the topic distribution over all words for topic $k$, and $\boldsymbol{A}_{w,k}$ is the probability of observing type $w$ given topic $k$.

**Anchors: Topic Representatives** The **anchor** method (Arora et al., 2012a) is based on the separability assumption (Donoho and Stodden, 2003), which assumes that each topic contains at least one namesake "anchor word" that has non-zero probability only in that topic. Intuitively, this means that each topic has unique, specific word that, when used, identifies that topic. For example, while "run", "base", "fly", and "shortstop" are associated with a topic about baseball, only "shortstop" is unambiguous, so it could serve as this topic's anchor word.

Let's assume that we knew what the anchor words were: a set $\mathcal{S}$ that indexes rows in $\boldsymbol{Q}$. Now consider the **conditional distribution** of word $i$, the probability of the rest of the vocabulary given an observation of word $i$; we represent this as $\bar{\boldsymbol{Q}}_{i,\cdot}$, as we can construct this by normalizing the rows of $\boldsymbol{Q}$. For an anchor word $s_a \in \mathcal{S}$, this will look like a topic; $\bar{\boldsymbol{Q}}_{\text{"shortstop"},\cdot}$ will have high probability for words associated with baseball.

The key insight of the **anchor** algorithm is that the conditional distribution of polysemous non-anchor words can be reconstructed as a linear combination of the conditional distributions of anchor words. For example, $\bar{\boldsymbol{Q}}_{\text{"fly"},\cdot}$ could be reconstructed by combining the anchor words "insecta", "boeing", and "shortshop". We represent the coefficients of this reconstruction as a matrix $\boldsymbol{C}$, where $C_{i,k} = p(z = k \mid w = i)$. Thus, for any word $i$,

$$\bar{\boldsymbol{Q}}_{i,\cdot} \approx \sum_{s_k \in \mathcal{S}} C_{i,k} \bar{\boldsymbol{Q}}_{s_k,\cdot} \tag{1}$$

The coefficient matrix is **not** the usual output of a topic modeling algorithm. The usual output is the probability of a word *given a topic*. The coefficient matrix $C$ is the probability of a topic *given a word*. We use Bayes rule to recover the topic distribution $p(w = i|z = k) \equiv$

$$\begin{aligned} A_{i,k} &\propto p(z = k|w = i)p(w = i) \\ &= C_{i,k} \sum_j \bar{Q}_{i,j} \end{aligned} \tag{2}$$

where $p(w)$ is the normalizer of $\boldsymbol{Q}$ to obtain $\bar{\boldsymbol{Q}}_{w,\cdot}$.

The geometric argument for finding the anchor words is one of the key contributions of Arora et al. (2012a) and is beyond the scope of this paper. The algorithms in Section 3 use the anchor selection subroutine unchanged. The difference in our approach is in how we discover the anchor coefficients $C$.

**From Anchors to Topics** After we have the anchor words, we need to find the coefficients that

best reconstruct the data $\bar{Q}$ (Equation 1). Arora et al. (2012a) chose the $C$ that minimizes the KL divergence between $\bar{Q}_{i,\cdot}$ and the reconstruction based on the anchor word's conditional word vectors $\sum_{s_k \in \mathcal{S}} C_{i,k} \bar{\boldsymbol{Q}}_{s_k,\cdot}$,

$$\boldsymbol{C}_{i,\cdot} = \text{argmin}_{\boldsymbol{C}_{i,\cdot}} D_{\text{KL}} \left( \bar{\boldsymbol{Q}}_{i,\cdot} \,||\, \sum_{s_k \in S} C_{i,k} \bar{\boldsymbol{Q}}_{s_k,\cdot} \right). \tag{3}$$

The **anchor** method is fast, as it only depends on the size of the vocabulary once the co-occurrence statistics $\boldsymbol{Q}$ are obtained. However, it does not support rich priors for topic models, while MCMC (Griffiths and Steyvers, 2004) and variational EM (Blei et al., 2003) methods can. This prevents models from using priors to guide the models to discover particular themes (Zhai et al., 2012), or to encourage sparsity in the models (Yao et al., 2009). In the rest of this paper, we correct this lacuna by adding regularization inspired by Bayesian priors to the **anchor** algorithm.

## 3 Adding Regularization

In this section, we add regularizers to the **anchor** objective (Equation 3). In this section, we briefly review regularizers and then add two regularizers, inspired by Gaussian ($L_2$, Section 3.1) and Dirichlet priors (Beta, Section 3.2), to the **anchor** objective function (Equation 3).

Regularization terms are ubiquitous. They typically appear as an additional term in an optimization problem. Instead of optimizing a function just of the data $x$ and parameters $\beta$, $f(x, \beta)$, one optimizes an objective function that includes a regularizer that is only a function of parameters: $f(w, \beta) + r(\beta)$. Regularizers are critical in staid methods like linear regression (Ng, 2004), in workhorse methods such as maximum entropy modeling (Dudík et al., 2004), and also in emerging fields such as deep learning (Wager et al., 2013).

In addition to being useful, regularization terms are appealing theoretically because they often correspond to probabilistic interpretations of parameters. For example, if we are seeking the MLE of a probabilistic model parameterized by $\beta$, $p(x|\beta)$, adding a regularization term $r(\beta) = \sum_{i=1}^{L} \beta_i^2$ corresponds to adding a Gaussian prior

$$f(\beta_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \text{exp} \left\{ -\frac{\beta_i^2}{2\sigma^2} \right\} \tag{4}$$

| Corpus | Train | Dev | Test | Vocab |
|---|---|---|---|---|
| NIPS | 1231 | 247 | 262 | 12182 |
| 20NEWS | 11243 | 3760 | 3726 | 81604 |
| NYT | 9255 | 2012 | 1959 | 34940 |

Table 2: The number of documents in the train, development, and test folds in our three datasets.

and maximizing log probability of the posterior (ignoring constant terms) (Rennie, 2003).

### 3.1 $L_2$ Regularization

The simplest form of regularization we can add is $L_2$ regularization. This is similar to assuming that probability of a word given a topic comes from a Gaussian distribution. While the distribution over topics is typically Dirichlet, Dirichlet distributions have been replaced by logistic normals in topic modeling applications (Blei and Lafferty, 2005) and for probabilistic grammars of language (Cohen and Smith, 2009).

Augmenting the **anchor** objective with an $L_2$ penalty yields

$$\begin{aligned} \boldsymbol{C}_{i,\cdot} = &\text{argmin}_{C_{i,\cdot}} D_{\text{KL}} \left( \bar{\boldsymbol{Q}}_{i,\cdot} \,||\, \sum_{s_k \in S} C_{i,k} \bar{\boldsymbol{Q}}_{s_k,\cdot} \right) \\ &+ \lambda \| \boldsymbol{C}_{i,\cdot} - \mu_{i,\cdot} \|_2^2, \end{aligned} \tag{5}$$

where regularization weight $\lambda$ balances the importance of a high-fidelity reconstruction against the regularization, which encourages the anchor coefficients to be close to the vector $\mu$. When the mean vector $\mu$ is zero, this encourages the topic coefficients to be zero. In Section 4.3, we use a non-zero mean $\mu$ to encode an informed prior to encourage topics to discover specific concepts.

### 3.2 Beta Regularization

The more common prior for topic models is a Dirichlet prior (Minka, 2000). However, we cannot apply this directly because the optimization is done on a row-by-row basis of the anchor coefficient matrix $\boldsymbol{C}$, optimizing $\boldsymbol{C}$ for a fixed word $w$ for and all topics. If we want to model the probability of a word, it must be the probability of word $w$ in a topic versus all other words.

Modeling this dichotomy (one versus all others in a topic) is possible. The constructive definition of the Dirichlet distribution (Sethuraman, 1994) states that if one has a $V$-dimensional multinomial $\theta \sim \text{Dir}(\alpha_1 \ldots \alpha_V)$, then the marginal distribution

of $\theta_w$ follows $\theta_w \sim \text{Beta}(\alpha_w, \sum_{i \neq w} \alpha_i)$. This is the tool we need to consider the distribution of a single word's probability.

This requires including the topic matrix as part of the objective function. The topic matrix is a linear transformation of the coefficient matrix (Equation 2). The objective for beta regularization becomes

$$\boldsymbol{C}_{i,\cdot} = \text{argmin}_{C_{i,\cdot}} D_{\text{KL}} \left( \bar{\boldsymbol{Q}}_{i,\cdot} \,||\, \sum_{s_k \in S} C_{i,k} \bar{\boldsymbol{Q}}_{s_k,\cdot} \right) - \lambda \sum_{s_k \in S} \log \left( \text{Beta}(A_{i,k}; a, b) \right), \quad (6)$$

where $\lambda$ again balances reconstruction against the regularization. To ensure the tractability of this algorithm, we enforce a convex regularization function, which requires that $a > 1$ and $b > 1$. If we enforce a uniform prior—$\mathbb{E}_{\text{Beta}(a,b)}[A_{i,k}] = \frac{1}{V}$—and and that the *mode* of the distribution is also $\frac{1}{V}$,[1] this gives us the following parametric form for $a$ and $b$:

$$a = \frac{x}{V} + 1, \text{ and } b = \frac{(V-1)x}{V} + 1 \quad (7)$$

for real $x$ greater than zero.

### 3.3 Initialization and Convergence

Equation 5 and Equation 6 are optimized using L-BFGS gradient optimization (Galassi et al., 2003). We initialize $\boldsymbol{C}$ randomly from $\text{Dir}(\alpha)$ with $\alpha = \frac{60}{V}$ (Wallach et al., 2009). We update $\boldsymbol{C}$ after optimizing all $V$ rows. The newly updated $\boldsymbol{C}$ replaces the old topic coefficients. We track how much the topic coefficients $\boldsymbol{C}$ change between two consecutive iterations $i$ and $i+1$ and represent it as $\Delta \boldsymbol{C} \equiv \|\boldsymbol{C}^{i+1} - \boldsymbol{C}^{i}\|_2$. We stop optimization when $\Delta \boldsymbol{C} \leq \delta$. When $\delta = 0.1$, the $L_2$ and unregularized anchor algorithm converges after a single iteration, while beta regularization typically converges after fewer than ten iterations (Figure 4).

## 4 Regularization Improves Topic Models

In this section, we measure the performance of our proposed regularized anchor word algorithms. We will refer to specific algorithms in bold. For example, the original anchor algorithm is **anchor**. Our $L_2$ regularized variant is **anchor-$L_2$**, and our beta regularized variant is **anchor-beta**. To provide conventional baselines, we also compare our methods against topic models from variational inference (Blei et al., 2003, **variational**) and MCMC (Griffiths and Steyvers, 2004; McCallum, 2002, **MCMC**).

We apply these inference strategies on three diverse corpora: scientific articles from the Neural Information Processing Society (Roweis, 2002, NIPS), Internet newsgroups postings (Lang, 2007, 20NEWS), and New York Times editorials (Sandhaus, 2008, NYT). Statistics for the datasets are summarized in Table 2. We split each dataset into a training fold (70%), development fold (15%), and a test fold (15%): the training data are used to fit models; the development set are used to select parameters (anchor threshold $M$, document prior $\alpha$, regularization weight $\lambda$); and final results are reported on the test fold.

We use two evaluation measures, held-out likelihood (Blei et al., 2003, **HL**) and topic interpretability (Chang et al., 2009; Newman et al., 2010, **TI**). Held-out likelihood measures how well the model can reconstruct held-out documents that the model has never seen before. This is the typical evaluation for probabilistic models. Topic interpretability is a more recent metric to capture how useful the topics can be to human users attempting to make sense of a large datasets.

Held-out likelihood cannot be computed with existing **anchor** algorithms, so we use the topic distributions learned from **anchor** as input to a reference variational inference implementation (Blei et al., 2003) to compute **HL**. This requires an additional parameter, the Dirichlet prior $\alpha$ for the per-document distribution over topics. We select $\alpha$ using grid search on the development set.

To compute **TI** and evaluate topic coherence, we use normalized pairwise mutual information (NPMI) (Lau et al., 2014) over topics' twenty most probable words. Topic coherence is computed against the NPMI of a reference corpus. For coherence evaluations, we use both intrinsic and extrinsic text collections to compute NPMI. Intrinsic coherence (TI-i) is computed on training and development data at development time and on training and test data at test time. Extrinsic coherence (TI-e) is computed from English Wikipedia articles, with disjoint halves (1.1 million pages each) used for distinct development and testing computations of TI-e.

[1] For $a, b < 1$, the expected value is still the uniform distribution but the mode lies at the boundaries of the simplex. This corresponds to a sparse Dirichlet distribution, which our optimization cannot at present model.

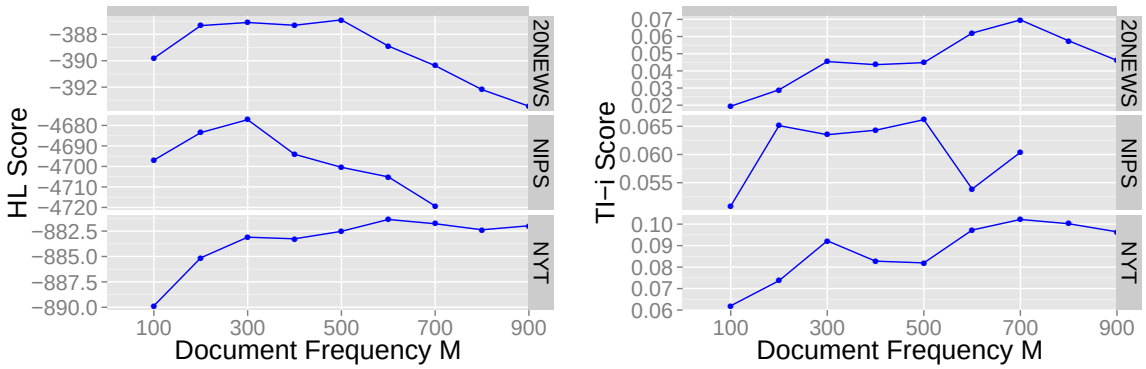


Figure 1: Grid search for document frequency $M$ for our datasets with 20 topics (other configurations not shown) on development data. The performance on both HL and TI score indicate that the unregularized **anchor** algorithm is very sensitive to $M$. The $M$ selected here is applied to subsequent models.

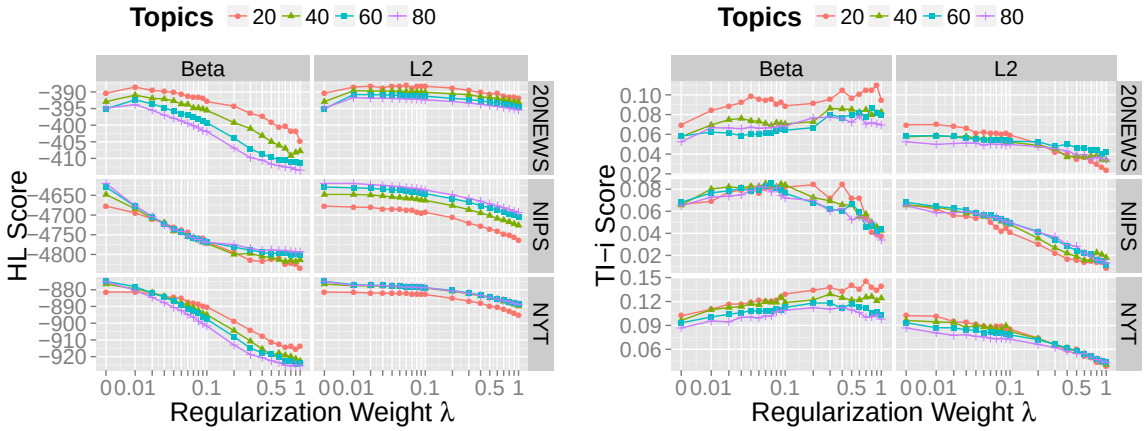


Figure 2: Selection of $\lambda$ based on HL and TI scores on the development set. The value of $\lambda = 0$ is equivalent to the original **anchor** algorithm; regularized versions find better solutions as the regularization weight $\lambda$ becomes non-zero.

### 4.1 Grid Search for Parameters on Development Set

**Anchor Threshold** A good anchor word must have a unique, specific context but also explain other words well. A word that appears only once will have a very specific cooccurence pattern but will explain other words' coocurrence poorly because the observations are so sparse. As discussed in Section 2, the **anchor** method uses document frequency $M$ as a threshold to only consider words with robust counts.

Because all regularizations benefit equally from higher-quality anchor words, we use cross-validation to select the document frequency cutoff $M$ using the unregularized **anchor** algorithm. Figure 1 shows the performance of **anchor** with different $M$ on our three datasets with 20 topics for our two measures HL and TI-i.

**Regularization Weight** Once we select a cutoff $M$ for each combination of dataset, number of topics $K$ and a evaluation measure, we select a regularization weight $\lambda$ on the development set. Figure 2 shows that **beta** regularization framework improves topic interpretability TI-i on all datasets and improved the held-out likelihood HL on 20NEWS. The $L_2$ regularization also improves held-out likelihood HL for the 20NEWS corpus (Figure 2).

In the interests of space, we do not show the figures for selecting $M$ and $\lambda$ using TI-e, which is similar to TI-i: **anchor-beta** improves the TI-e score on all datasets, **anchor-**$L_2$ improves TI-e on 20NEWS and NIPS with 20 topics and NYT with 40 topics.

### 4.2 Evaluating Regularization

With document frequency $M$ and regularization weight $\lambda$ selected from the development set, we compare the performance of those models on the test set. We also compare with standard implementations of Latent Dirichlet Allocation: Blei's LDAC (**variational**) and Mallet (**mcmc**). We run 100 iterations for LDAC and 5000 iterations for Mallet.

Each result is averaged over three random runs and appears in Figure 3. The highly-tuned, widely-used implementations uniformly have better held-out likelihood than **anchor**-based methods, but the much faster **anchor** methods are often comparable. Within **anchor**-based methods, $L_2$-regularization offers comparable held-out likelihood as unregularized **anchor**, while **anchor-beta** often has better interpretability. Because of the mismatch between the specialized vocabulary of NIPS and the general-purpose language of Wikipedia, TI-e has a high variance.

### 4.3 Informed Regularization

A frequent use of priors is to add information to a model. This is not possible with the existing **anchor** method. An informed prior for topic models seeds a topic with words that describe a topic of interest. In a topic model, these seeds will serve as a "magnet", attracting similar words to the topic (Zhai et al., 2012).

We can achieve a similar goal with **anchor-**$L_2$. Instead of encouraging anchor coefficients to be zero in Equation 5, we can instead encourage word probabilities to close to an arbitrary mean $\mu_{i,k}$. This vector can reflect expert knowledge.

One example of a source of expert knowledge is Linguistic Inquiry and Word Count (Pennebaker and Francis, 1999, LIWC), a dictionary of keywords related to sixty-eight psychological concepts such as positive emotions, negative emotions, and death. For example, it associates "excessive, estate, money, cheap, expensive, living, profit, live, rich, income, poor, etc." for the concept materialism.

We associate each anchor word with its closest LIWC category based on the cooccurrence matrix $\boldsymbol{Q}$. This is computed by greedily finding the anchor word that has the highest cooccurrence score for any LIWC category: we define the score of a category to anchor word $w_{s_k}$ as $\sum_i Q_{s_k,i}$, where $i$ ranges over words in this category; we compute the scores of all categories to all anchor words; then we find the highest score and assign the category to that anchor word; we greedily repeat this process until all anchor words have a category.

Given these associations, we create a goal mean $\mu_{i,k}$. If there are $L_i$ anchor words associated with LIWC word $i$, $\mu_{i,k} = \frac{1}{L_i}$ if this keyword $i$ is associated with anchor word $w_{s_k}$ and zero otherwise.

We apply **anchor-**$L_2$ with informed priors on NYT with twenty topics and compared the topics against the original topics from **anchor**. Table 3 shows that the topic with anchor word "soviet", when combined with LIWC, draws in the new words "bush" and "nuclear"; reflecting the threats of force during the cold war. For the topic with topic word "arms", when associated with the LIWC category with the terms "agree" and "agreement", draws in "clinton", who represented a more conciliatory foreign policy compared to his republican predecessors.

## 5 Discussion

Having shown that regularization can improve the **anchor** topic modeling algorithm, in this section we discuss *why* these regularizations can improve the model and the implications for practitioners.

**Efficiency** Efficiency is a function of the number of iterations and the cost of each iteration. Both **anchor** and **anchor-**$L_2$ require a single iteration, although the latter's iteration is slightly more expensive. For **beta**, as described in Section 3.2, we update anchor coefficients $C$ row by row, and then repeat the process over several iterations until it converges. However, it often converges within ten iterations (Figure 4) on all three datasets: this requires much fewer iterations than MCMC or variational inference, and the iterations are less expensive. In addition, since we optimize each row $C_{i,\cdot}$ independently, the algorithm can be easily parallelized.

**Sensitivity to Document Frequency** While the original **anchor** is sensitive to the document frequency $M$ (Figure 1), adding regularization makes this less critical. Both **anchor-**$L_2$ and **anchor-beta** are less sensitive to $M$ than **anchor**.

To highlight this, we compare the topics of **anchor** and **anchor-beta** when $M = 100$. As Table 4 shows, the words "article", "write", "don" and "doe" appear in most of **anchor**'s topics. While **anchor-**$L_2$ also has some bad topics, it still can find reasonable topics, demonstrating **anchor-beta**'s greater robustness to suboptimal $M$.

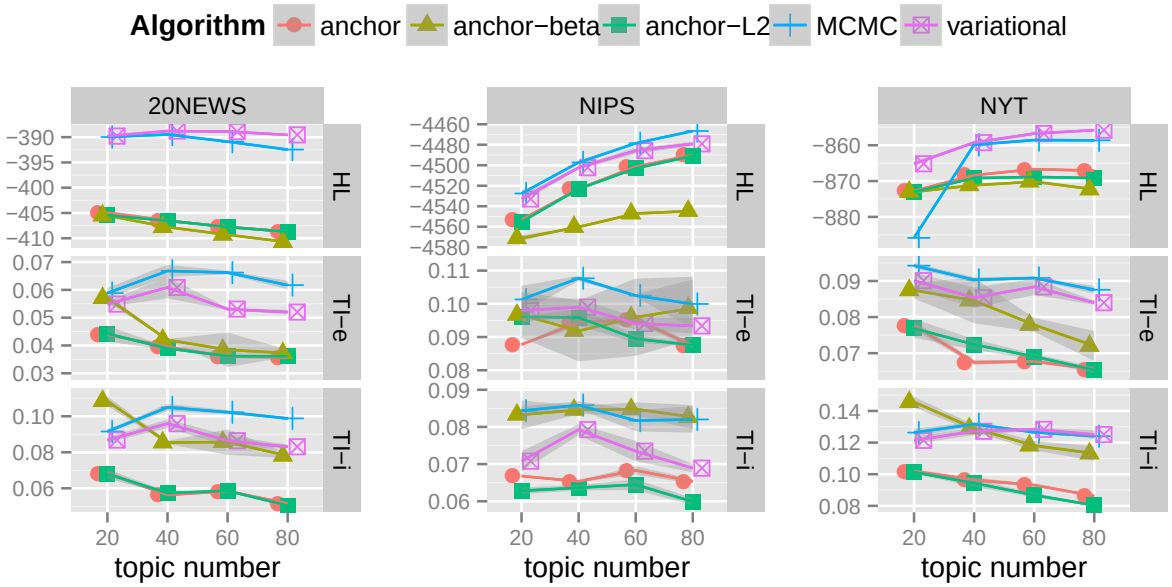


Figure 3: Comparing **anchor-beta** and **anchor-**$L_2$ against the original **anchor** and the traditional **variational** and **MCMC** on HL score and TI score. **variational** and **mcmc** provide the best held-out generalization. **anchor-beta** sometimes gives the best TI score and is consistently better than **anchor**. The specialized vocabulary of NIPS causes high variance for the extrinsic interpretability evaluation (TI-e).

| Topic | Shared Words | Original (Top, green) vs. Informed $L_2$ (Bottom, orange) |
|---|---|---|
| soviet | american make president **soviet** union **war** years | gorbachev moscow russian force economic world europe political communist lead reform germany country<br>**military** state **service** washington **bush** **army** unite **chief troops officer** **nuclear** time week |
| district | **assembly** board city **county district member** state york | representative manhattan brooklyn queens election bronx council island local incumbent housing municipal<br>**people party group social** **republican** year make years **friend** **vote** **compromise** million |
| peace | american force government israel **peace** political president state unite washington | war military country minister leaders nation world palestinian israeli election<br>**offer justice aid deserve** make **bush** years **fair** **clinton** **hand** |
| arms | **arms** bush congress force iraq make north nuclear president state washington weapon | administration treaty missile defense war military korea reagan<br>**agree agreement** american **accept** unite **share** **clinton** years |
| trade | **administration** america american country **economic** government make president state **trade** unite washington | world market japan foreign china policy price political<br>**business economy** **congress** year years **clinton** **bush** **buy** |

Table 3: Examples of topic comparison between **anchor** and informed **anchor-**$L_2$. A topic is labeled with the anchor word for that topic. The **bold** words are the informed prior from LIWC. With an informed prior, relevant words appear in the top words of a topic; this also draws in other related terms (red).

**$L_2$ (Sometimes) Improves Generalization** As Figure 2 shows, **anchor-**$L_2$ can sometimes improve the held-out likelihood on the development set on the smaller 20NEWS corpus. However, the $\lambda$ selected on development data does not always improve test set performance. This, in Figure 3, **anchor-beta** closely tracks **anchor**. Thus, $L_2$ regularization does not hurt generalization while imparting benefits for expressiveness and robustness to parameter settings.

**Beta Improves Interpretability** Figure 3 shows that **anchor-beta** improves topic interpretability (TI) compared to unregularized anchor methods. In this section, we try to understand why.

We first compare the topics from the original **anchor** against **anchor-beta** to analyze the topics qualitatively. Table 5 shows that **beta** regularization promotes rarer words within a topic and de-

| Topic | anchor | anchor-beta |
|---|---|---|
| frequently | article write don doe make time people good file question | article write don doe make people time good email file |
| debate | write article people make don doe god key government time | people make god article write don doe key point government |
| wings | game team write wings article win red play hockey year | game team wings win red hockey play season player fan |
| stats | player team write game article stats year good play doe | stats player season league baseball fan team individual playoff nhl |
| compile | program file write email doe windows call problem run don | compile program code file ftp advance package error windows sun |

Table 4: Topics from **anchor** and **anchor-beta** with $M = 100$ on 20NEWS with 20 topics. Each topic is identified with its associated anchor word. When $M = 100$, the topics of **anchor** suffer: the four colored words appear in almost every topic. **anchor-beta**, in contrast, is less sensitive to suboptimal $M$.

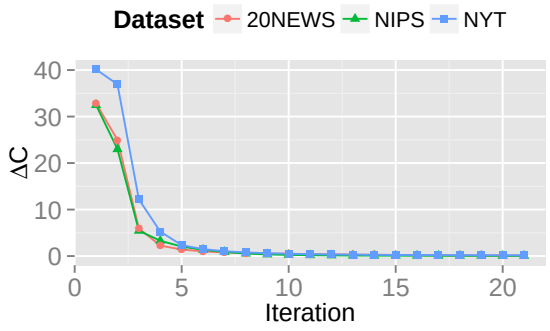


Figure 4: Convergence of anchor coefficient $C$ for **anchor-beta**. $\Delta C$ is the difference of current $C$ from the $C$ at the previous iteration. $C$ is converged within ten iterations for all three datasets.

motes common words. For example, in the topic about hockey with the anchor word game, "run" and "good"—ambiguous, polysemous words—in the unregularized topic are replaced by "playoff" and "trade" in the regularized topic. These words are less ambiguous and more likely to make sense to a consumer of topic models.

Figure 5 shows why this happens. Compared to the unregularized topics from **anchor**, the beta regularized topic steals from the rich and creates a more uniform distribution. Thus, highly frequent words do not as easily climb to the top of the distribution, and the topics reflect topical, relevant words rather than globally frequent terms.

## 6 Conclusion

A topic model is a popular tool for quickly getting the gist of large corpora. However, running such an analysis on these large corpora entail a substantial computational cost. While techniques such as **anchor** algorithms offer faster solutions, it comes at the cost of the expressive priors common in Bayesian formulations.

This paper introduces two different regularizations that offer users more interpretable models and the ability to inject prior knowledge without sacrificing the speed and generalizability of the underlying approach. However, one sacrifice that this approach does make is the beautiful theoretical guarantees of previous work. An important piece of future work is a theoretical understanding of generalizability in extensible, regularized models.

Incorporating other regularizations could further improve performance or unlock new applications. Our regularizations do not explicitly encourage sparsity; applying other regularizations such as $L_1$ could encourage true sparsity (Tibshirani, 1994), and structured priors (Andrzejewski et al., 2009) could provide efficient incorporation of constraints on topic models.

These regularizations could also be applied to other spectral algorithms for latent variables models, improving the performance for other NLP tasks such as latent variable PCFGs (Cohen et al., 2013) and HMMs (Anandkumar et al., 2012), combining the flexibility and robustness offered by priors with the speed and accuracy of new, scalable algorithms.

## Acknowledgments

We would like to thank the anonymous reviewers, Hal Daumé III, Ke Wu, and Ke Zhai for their helpful comments. This work was supported by NSF Grant IIS-1320538. Boyd-Graber is also supported by NSF Grant CCF-1018625. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily

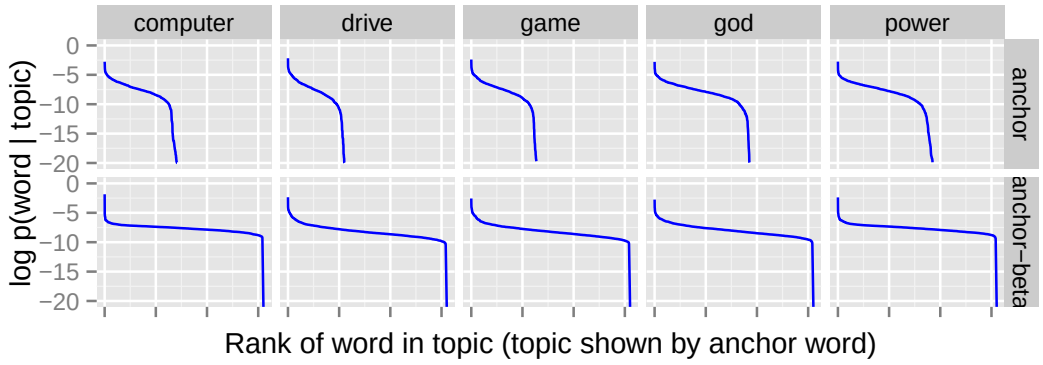


Figure 5: How beta regularization influences the topic distribution. Each topic is identified with its associated anchor word. Compared to the unregularized **anchor** method, **anchor-beta** steals probability mass from the "rich" and prefers a smoother distribution of probability mass. These words often tend to be unimportant, polysemous words common across topics.

| Topic | Shared Words | **anchor** (Top, green) vs. **anchor-beta** (Bottom, orange) |
|---|---|---|
| computer | computer means science screen | system phone university problem doe work windows internet software chip mac set fax technology information data |
| | | quote mhz pro processor ship remote print devices complex cpu electrical transfer ray engineering serial reduce |
| power | power play period supply ground light battery engine | car good make high problem work back turn control current small time |
| | | circuit oil wire unit water heat hot ranger input total joe plug |
| god | god jesus christian bible faith church life christ belief religion hell word lord truth love | people make things true doe |
| | | sin christianity atheist peace heaven |
| game | game team player play win fan hockey season baseball red wings score division league goal leaf cup toronto | run good |
| | | playoff trade |
| drive | drive disk hard scsi controller card floppy ide mac bus speed monitor switch apple cable internal port meg | problem work |
| | | ram pin |

Table 5: Comparing topics—labeled by their anchor word—from **anchor** and **anchor-beta**. With beta regularization, relevant words are promoted, while more general words are suppressed, improving topic coherence.

reflect the view of the sponsor.

## References

Animashree Anandkumar, Daniel Hsu, and Sham M. Kakade. 2012. A method of moments for mixture models and hidden markov models. In *Proceedings of Conference on Learning Theory*.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the International Conference of Machine Learning*.

Sanjeev Arora, Rong Ge, Yoni Halpern, David M. Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2012a. A practical algorithm for topic modeling with provable guarantees. *CoRR*, abs/1212.4777.

Sanjeev Arora, Rong Ge, and Ankur Moitra. 2012b. Learning topic models - going beyond svd. *CoRR*, abs/1204.1956.

David M. Blei and John D. Lafferty. 2005. Correlated topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, Athens, Greece.

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Shay B. Cohen and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

S. B. Cohen, K. Stratos, M. Collins, D. P. Foster, and L. Ungar. 2013. Experiments with spectral learning of latent-variable PCFGs. In *Proceedings of NAACL*.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the Association for Computational Linguistics*.

David Donoho and Victoria Stodden. 2003. When does nonnegative matrix factorization give correct decomposition into parts? page 2004. MIT Press.

Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. 2004. Performance guarantees for regularized maximum entropy density estimation. In *Proceedings of Conference on Learning Theory*.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the Association for Computational Linguistics*.

Jenny Rose Finkel and Christopher D. Manning. 2009. Hierarchical bayesian domain adaptation. In *Conference of the North American Chapter of the Association for Computational Linguistics*, Morristown, NJ, USA.

Mark Galassi, Jim Davies, James Theiler, Brian Gough, Gerard Jungman, Michael Booth, and Fabrice Rossi. 2003. *Gnu Scientific Library: Reference Manual*. Network Theory Ltd.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228–5235.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of Emperical Methods in Natural Language Processing*.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2013. Interactive topic modeling. *Machine Learning Journal*.

Matt L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. University of Illinois Press.

Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. 2005. The spectral method for general mixture models. In *Proceedings of Conference on Learning Theory*.

Ken Lang. 2007. 20 newsgroups data set.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

Thomas P. Minka. 2000. Estimating a dirichlet distribution. Technical report, Microsoft. http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Andrew Y. Ng. 2004. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the International Conference of Machine Learning*.

Michael Paul and Roxana Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Association for the Advancement of Artificial Intelligence*.

James W. Pennebaker and Martha E. Francis. 1999. *Linguistic Inquiry and Word Count*. Lawrence Erlbaum, 1 edition, August.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Jason Rennie. 2003. On l2-norm regularization and the Gaussian prior.

Sam Roweis. 2002. NIPS 1-12 Dataset.

Evan Sandhaus. 2008. The New York Times annotated corpus. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19.

Jayaram Sethuraman. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Robert Tibshirani. 1994. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Stefan Wager, Sida Wang, and Percy Liang. 2013. Dropout training as adaptive regularization. In *Proceedings of Advances in Neural Information Processing Systems*, pages 351–359.

Hanna Wallach, David Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Proceedings of Advances in Neural Information Processing Systems*.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Knowledge Discovery and Data Mining*.

Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad Alkhouja. 2012. Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of World Wide Web Conference*.