Evaluating Regularized Anchor Words

Thang Nguyen
iSchool
University of Maryland
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

We perform a comprehensive examination of the recently proposed anchor method for topic model inference using topic interpretability and held-out likelihood measures. After measuring the sensitivity to the anchor selection process, we incorporate L_2 and Beta regularization into the optimization objective in the recovery step. Preliminary results show that L_2 improves heldout likelihood, and Beta regularization improves topic interpretability.

1 Introduction

Topic models are unsupervised methods that learn thematic structures from a set of documents. These models explain documents' content as an admixture over topics, the namesake distributions over the vocabulary that explain a dataset's primary themes. Given a collection of documents, the fundamental problem of topic models is to discover the topics and document allocations that best explain a corpus. Typical solutions use MCMC [1] or variational EM [2].

Recently, however, new solutions provide provable polynomial-time alternatives. Arora et al. [3] present a non-negative matrix factorization technique which assumes that the data evince "anchor words" which can separate each topic (henceforth called **anchor**); each topic contains at least one anchor word (a word which has non-zero probability only for that topic). Related techniques use spectral decomposition to match the moments of the assumed generating distribution [4]. Unlike search-based methods, which can be caught in local minima, these techniques are guaranteed to find global optima (henceforth called **svd**, as these techniques use eigenvalue decompositions).

However, these techniques are not a panacea to practitioners of topic models. First, they are not flexible enough to incorporate the rich priors that make Bayesian topic models so attractive [5]. In an ideal situation, each topic should reflect the co-occurring relationship between words and at the same time be distinct from each other to convey information. This situation suggests using symmetric prior over topic distributions [5]. In this paper, we propose regularized versions of **anchor** that provide many of the same advantages of rich Bayesian priors. Second, these new techniques from the theory community have not been evaluated using traditional topic model evaluations; to vet our modified algorithms, we also compare to **anchor** on held-out likelihood [2] and topic interpretability [6, 7].

2 Background

The **anchor** method in [8] is based on the separability assumption [9], which assumes that each topic contains at least one anchor word that has non-zero probability only in that topic. Thus this **anchor** method includes two steps: first, select anchor words for each topic, and then reconstruct the topic distributions based on the anchor words.

For both steps, select anchor words and recover topic distributions, **anchor** uses the word-word co-occurrence count matrix Q (of size $V \times V$, V is the vocabulary size) as the input [3]. Given unlimited documents, each element of matrix Q can be represented as the joint distribution of corresponding

words, $Q_{i,j} = p(w_1 = i, w_2 = j)$. Therefore, if we denote $\bar{Q}$ as the row-normalized of Q then each element of $\bar{Q}$ can be interpreted as the conditional probability $\bar{Q}_{i,j} = p(w_2 = j \mid w_1 = i)$ [8].

The first step of **anchor** is to find anchor words for each topic. Given the row-normalized word co-occurrence matrix $\bar{Q}$ and unlimited documents, the convex hull of the rows in $\bar{Q}$ will be a simplex where the vertices of this simplex correspond to the anchor words [8]. The authors of the **anchor** method suggest filtering candidates only to those which appear in at least M documents.

Then the topic recovery step uses these anchor words to recover the topics based on co-occurrence statistics of words with the anchor words. This is possible because any row of $\bar{Q}$ lies in the convex hull of the rows corresponding to the anchor words. Thus for each row $\bar{Q}_{i,\cdot}$, [8] tries to find the optimal coefficients of the anchor words to minimize the KL divergence to that row. Then the topic distributions over words can be recovered based on the coefficients matrix.

The **anchor** method is fast, as it only depends on the size of the vocabulary once the co-occurrence statistics Q are obtained. However, it does not support rich priors for topic models, while the MCMC [1] or variational EM [2] based methods can. This prevents models from using priors to guide the models to discover particular themes [10], or to encourage sparsity in the models [11]. In the rest of this paper, we investigate regularization to **anchor** to incorporate these rich priors.

3 Adding Regularization

While the original **anchor** method doesn't include regularization, in this section, we augment the objective function of the topic recovery step to include penalties that have the same functional form as Bayesian priors. The unaugmented **anchor** objective function is to find topics $C_{i,k}$, where $C_{i,k}$ is the probability of topic k given word i (the reverse of the typical topic model formulation) such that

$$C_{i\cdot} = argmin_{\vec{C_i}} D_{KL}\left(\bar{Q}_i || \sum_{s_k \in S} C_{i,k}\bar{Q}_{s_k}\right), \tag{1}$$

where $S = s_1, s_2, \cdots, s_K$ are the indices of the anchor words, and $\bar{Q}(s_k)$ as the conditional co-occurrence probability of anchor words for topic k. This objective function minimizes the KL divergence between $\bar{Q}$ and a linear combination of the rows with anchor words. This views the topic distribution C as a free multinomial parameter. In this section, we add additional constraints on C for this objective function that correspond to two different priors: a Gaussian prior and a Dirichlet prior.

3.1 L_2 Regularization

Gaussian priors, equivalent to L_2 regularization, are one of the most common priors used in statistical modeling. We can add L_2 regularization to the reconstruction objective,

$$C_{i\cdot} = argmin_{\vec{C_i}} D_{KL}(\bar{Q}_i || \sum_{s_k \in S} C_{i,k}\bar{Q}_{s_k}) + \lambda \|C_i\|_2, \tag{2}$$

where λ balances the importance of a high-fidelity reconstruction against the regularization.

3.2 Beta Regularization

The Dirichlet prior over topics [2] encourages the topic sparsity. However, we cannot directly add a Dirichlet prior term because the **anchor** method's optimization considers the probability of a single word in *all* topics. However, because of marginal consistency of the Dirichlet [12], we can consider the probability of a single word in a topic (against the probability of all *other* words in a topic) as an appropriately parameterized Beta distribution.

Define $\beta_{k,i} = p(w = i|z = k), i \in V$ and $s_k \in S$, then the objective for beta regularization becomes:

$$C_{i\cdot} = argmin_{\vec{C_i}} D_{KL}(\bar{Q}_i \| \sum_{s_k \in S} C_{i,k}\bar{Q}_{s_k}) - \lambda \sum_{s_k \in S} \log\left(\text{Beta}(\beta_{k,i}; \alpha, (V-1)\alpha)\right), \tag{3}$$

where λ again balances reconstruction against the regularization.

In practice, we initialize C matrix from Dirichlet(α), and we select α following [5] as $\alpha = \frac{60}{V}$. Then we iteratively update C row by row, untill convergence.

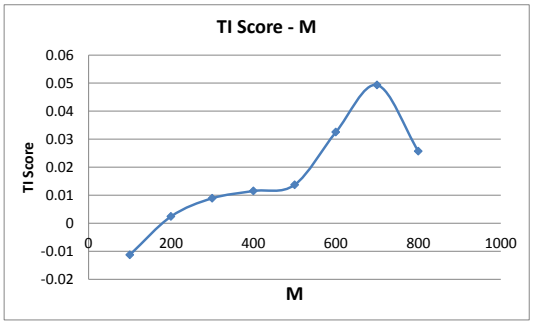

Figure 1: TI Score for word count thresholds M on 20 NEWS group dataset.

4 Experimental Results

We checked our proposed regularization methods on 20NewsGroups dataset. We use their default train and test split (11243 train documents and 7488 test documents), and use half of the test split as develop set, and the remaining as the test; and the vocabulary includes 81600 word types.

We use two evaluation measures, held-out likelihood (denoted as **HL**) [2] and topic interpretability (denoted as **TI**) [6, 7]. **HL** measures generalization, and **TI** measures how "interpretable" topics are. We use a reference variational inference implementation (**LDA-C**) [2] to compute **HL** given topic distributions (as **anchor** is undefined for test documents). To evaluate topic coherence, we use normalized pairwise mutual information (NPMI) [13] over top ten words extracted from each topic.

We select M and λ using grid search on development data and apply topics learned with those parameters on the test set. We compare anchor method with L_2 and beta regularization (denoted as **anchor-**L_2 and **anchor-beta** respectively) with the original anchor method (denoted as **anchor**). For each set of parameters on one algorithm, we run 5 times and average over all the scores.

Anchor selection Because of the sensitivity to M, the word count threshold, we include M as a parameter optimized by grid search (Figure 1). Qualitative inspection of the topics (see Appendix) confirms these results, from which, we can clearly see that it detected the "sports" topic, "computer" topic, "religion" topic, "government security" topic, etc. very clearly when $M = 700$.

Since this **anchor** algorithm is very sensitive to the extracted anchor words, we can consider better ways to extract the anchor word candidates rather than document frequency, for example, tf-idf, etc.

Regularization and Evaluation For a given word count cut-off, we can examine the trend of our evaluation measures for different values of λ and number of topics K on our develop set (Figure 2). For larger number of topics ($K > 20$), L_2 regularization improves held-out prediction, and Beta regularization improves coherence (**TI**). However, when $K = 20$, neither of the two regularization are better than **anchor**; we hypothesize this is because there is sufficient data for each topic to learn effective topics.

Given the parameters selected on development data, we apply these parameters to test data (Table 1). These obtained comparable results; word-topic profiles are available in the Appendix 6.3 and 6.4.

5 Conclusion

This paper introduces two different regularizations to spectral learning methods of topic modeling, and evaluates the resulting topics using coherence and held-out likelihood. While the regularization did not show much effect when the topic number is small, when the topic number became larger, L_2 improves heldout likelihood score, and Beta regularization improves coherence.

We are investigating other regularizations such as L_1 regularization, using adaptive settings of λ, and initializing the regularized models with the results of unconstrained inference.

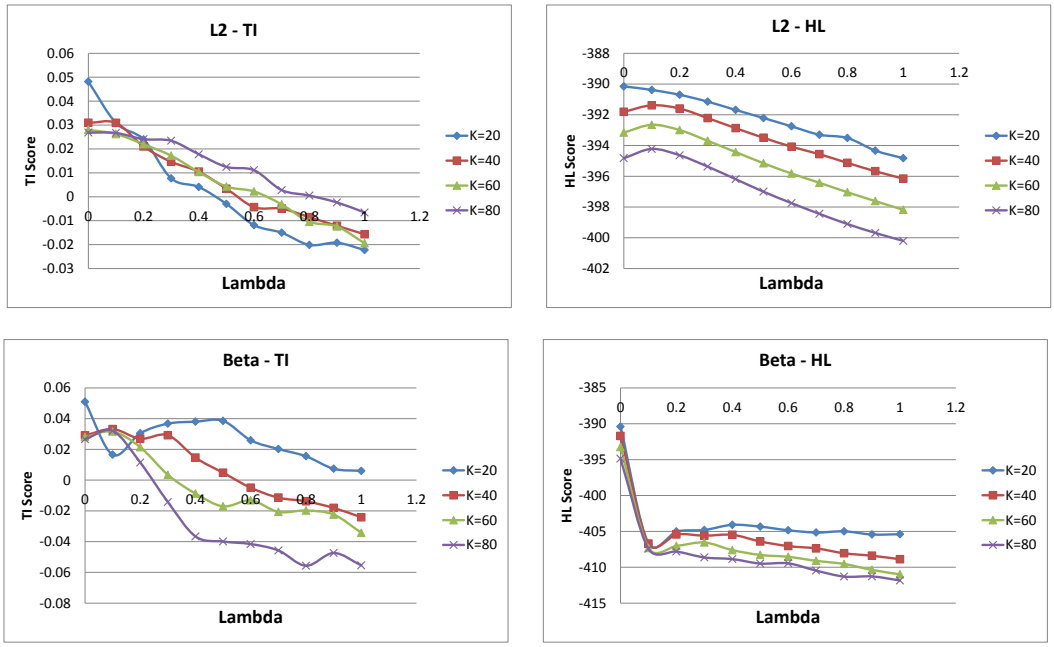

Figure 2: TI Score and held-out likelihood score for L_2 regularization on 20news dataset, M = 700.

	↑TI				↑HL			
	K=20	K=40	K=60	K=80	K=20	K=40	K=60	K=80
anchor	**0.0499**	0.0314	0.0277	0.0255	**-407.4**	-408.8	-410.5	-411.9
anchor-L_2	0.0318	0.0273	0.0255	0.0256	-407.5	**-408.0**	**-409.8**	**-411.3**
anchor-beta	0.0152	**0.0326**	**0.0322**	**0.0321**	-424.4	-424.6	-425.0	-425.2

Table 1: Apply the selected parameters on the test data of 20 NEWS. **anchor-beta** obtained better results on **TI** score, while **anchor-L_2** got slightly better results on **HL** score.

References

[1] Griffiths, T. L., M. Steyvers. Finding scientific topics. *PNAS*, 101(Suppl 1):5228–5235, 2004.

[2] Blei, D. M., A. Ng, M. Jordan. Latent Dirichlet allocation. *JMLR*, 2003.

[3] Arora, S., R. Ge, A. Moitra. Learning topic models - going beyond svd. *CoRR*, abs/1204.1956, 2012.

[4] Anandkumar, A., D. P. Foster, D. Hsu, et al. Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. *CoRR*, abs/1204.6703, 2012.

[5] Wallach, H., D. Mimno, A. McCallum. Rethinking LDA: Why priors matter. In *NIPS*. 2009.

[6] Chang, J., J. Boyd-Graber, C. Wang, et al. Reading tea leaves: How humans interpret topic models. In *NIPS*. 2009.

[7] Newman, D., J. H. Lau, K. Grieser, et al. Automatic evaluation of topic coherence. In *NAACL*. 2010.

[8] Arora, S., R. Ge, Y. Halpern, et al. A practical algorithm for topic modeling with provable guarantees. *CoRR*, abs/1212.4777, 2012.

[9] Donoho, D., V. Stodden. When does non-negative matrix factorization give correct decomposition into parts? page 2004. MIT Press, 2003.

[10] Zhai, K., J. Boyd-Graber, N. Asadi, et al. Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *WWW*. 2012.

[11] Yao, L., D. Mimno, A. McCallum. Efficient methods for topic model inference on streaming document collections. In *KDD*. 2009.

[12] Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[13] Bouma, G. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*. 2009.

6 Appendix

6.1 20news topics, K = 20, $\lambda = 0$, $\lambda = 0.1$, M = 100

Topic	Anchor word	Top 10 words
Topic 1	max	max brain cheer clipper ticket pgp electrical andy traffic att
Topic 2	van	van write win article doe team andrew april year email
Topic 3	frequently	article write don doe make time people good system question
Topic 4	debate	write article people make don doe key government time point
Topic 5	stats	player write team game article stats year good play don
Topic 6	danger	write article don people make system doe government problem good
Topic 7	ignorance	write article don people god doe make post ignorance time
Topic 8	wings	game team write wings article win red play hockey year
Topic 9	mailing	email write list article doe address mailing internet mail people
Topic 10	eternal	god write article don jesus people christian make time bible
Topic 11	touch	write article don good make car time doe problem people
Topic 12	letters	write article don email doe case make people good letters
Topic 13	vga	drive card email doe windows monitor write sale system offer
Topic 14	geb	write article don people gordon banks geb make good doe
Topic 15	default	window write file problem article windows doe system make set
Topic 16	armenia	armenian write people turkish article armenia war government israel jew
Topic 17	compile	program write file email doe call windows problem don run
Topic 18	update	windows file write driver system doe version email update problem
Topic 19	period	write article period power play don game make good year
Topic 20	plenty	write article don make people good doe car time work

6.2 20news topics, K = 20, $\lambda = 0$, $\lambda = 0.1$, M = 700

Topic	Anchor word	Top 10 words
Topic 1	drive	drive disk hard scsi controller card problem floppy ide mac
Topic 2	god	god jesus christian people bible faith church life christ belief
Topic 3	game	game team player play win fan hockey season run baseball
Topic 4	file	file windows ftp driver dos version site image directory problem
Topic 5	article	article don people make time back good work isn doe
Topic 6	list	list mailing address add people send post book marc interest
Topic 7	program	program window call doe advance problem application run windows give
Topic 8	power	power car play period good supply make ground light battery
Topic 9	government	government people key state law make israel gun israeli encryption
Topic 10	part	part max doe end air make cut western call include
Topic 11	support	support doe driver card version mode video system information work
Topic 12	group	group post don question posting people read time newsgroup create
Topic 13	line	line problem window display set doe place point subject find
Topic 14	computer	computer system phone university problem doe science means work windows
Topic 15	year	year team good years player time win car make play
Topic 16	buy	buy car price good bike doe don sell make cheap
Topic 17	write	write don make people time good doe post back thing
Topic 18	number	number don key call phone doe question order chip company
Topic 19	john	john move internet receive doe black full jewish posting include
Topic 20	email	email sale offer send address fax interest advance internet mail

6.3 20news topics, M = 700, K = 40, $\lambda = 0$

Topic	Anchor word	Top 10 words
Topic 1	drive	drive disk hard scsi controller card floppy ide mac problem
Topic 2	god	god jesus christian people bible church christ life belief faith
Topic 3	game	game team player play win fan hockey season run baseball
Topic 4	file	file windows ftp driver dos version site image directory doe
Topic 5	list	list mailing address add people doe marc send user mike
Topic 6	article	article don people make bob back mark didn steve gordon
Topic 7	program	program window advance doe windows run application display object user
Topic 8	power	power car play period supply ground battery light high current
Topic 9	part	part max doe end air cut make western individual pay
Topic 10	support	support doe driver card version mode video cards graphics software
Topic 11	government	government people key encryption law armenian public clipper make system
Topic 12	line	line doe display subject place area screen find easy note
Topic 13	group	group don question create discussion newsgroup posting wrong tom news
Topic 14	computer	computer system phone means windows software problem mac quote screen
Topic 15	write	write don make people good doe didn lot doesn opinion
Topic 16	year	year team player years good win play car make big
Topic 17	buy	buy car price good bike doe cheap sell card dealer
Topic 18	email	email send address fax advance reply offer mail sale internet
Topic 19	john	john move receive doe internet full jewish posting black tom
Topic 20	order	order point don find question place net mail long give
Topic 21	times	times time good doe joe tire method advice place point
Topic 22	david	david guy time care internet koresh back don great netcomcom
Topic 23	put	put back don make man space people time bike face
Topic 24	number	number phone don key company question numbers chip answer read
Topic 25	bit	bit key data speed fast work bike time chip problem
Topic 26	hear	hear doug happen eat found patient slot problem true mark
Topic 27	claim	claim people evidence israel doe make fact objective truth israeli
Topic 28	information	information doe interest book data appreciated info find point source
Topic 29	change	change problem don system similar make work start things turn
Topic 30	university	university internet fax research science institute usa phone view department
Topic 31	include	include sale offer condition price sell cover original good shipping
Topic 32	real	real don test time posting people close work true doe
Topic 33	kind	kind don max people doe lot dan soul age send
Topic 34	post	post won posting faq read response doe message question final
Topic 35	isn	isn henry keith andy clipper don work assume doe people
Topic 36	bad	bad good don make time people problem experience things comment
Topic 37	show	show men world child people faith time found study james
Topic 38	set	set window problem work source command colors start doe color
Topic 39	call	call don give make good case peter reply love friend
Topic 40	state	state people bill law gun live country carry israel years

6.4 20news topics, M = 700, K = 40, $\lambda = 0.1$, anchor-beta

Topic	Anchor word	Top 10 words
Topic 1	drive	drive card software hard data mac disk machine monitor speed
Topic 2	god	god true reason life man christian person love exist assume
Topic 3	file	file windows advance driver version found source works check memory
Topic 4	game	game team play mike player home win fan guy season
Topic 5	government	key government pay chip law today issue public gun kill
Topic 6	list	list add white class complete texas cheer count mailing marc
Topic 7	article	article dave bob brain gordon doctor fit banks andrew foot
Topic 8	program	program window application simple manager function object algorithm values associate
Topic 9	power	power light special supply ground period unit washington battery circuit
Topic 10	part	part end max air cut mile individual begin parts ron
Topic 11	group	group posting deal tom newsgroup folks specific reader personally curious
Topic 12	support	support mode mouse cards release higher motif technical greg resolution
Topic 13	computer	computer phone means quote pro ray ship electronic processor corp
Topic 14	line	line signal draw noise wall distribution led newsgroups bottom drawing
Topic 15	buy	buy price bike current cost cheap thinking engine ride worth
Topic 16	write	write don make people good time system problem work question
Topic 17	number	number doe prefer professional charles exact procedure equivalent telephone comparison
Topic 18	year	year top city numbers early past smith san stay george
Topic 19	bit	bit suggest stick clock doe turning operations random jump rate
Topic 20	email	email internet send address info fax mail offer reply sell
Topic 21	john	john move receive jewish letters weren doe black cross brain
Topic 22	hear	hear food doug slot disease dog eat hours patient hall
Topic 23	order	order tony equipment piece cross finger bell warning technique doe
Topic 24	real	real test close vehicle treat planet doe art posting letters
Topic 25	times	times advice page tire pain heavy enter doe central length
Topic 26	change	change similar normal avoid hole oil review exact larger drink
Topic 27	university	university research technology science view usa radio school major department
Topic 28	david	david koresh stephen guy fly doe cambridge shoot stupid highly
Topic 29	put	put pull moon putting keeping doe conditions potential shut wheel
Topic 30	claim	claim evidence israel study taking objective frank water finally driving
Topic 31	information	information appreciated reference project chips development medical services andor surface
Topic 32	include	include cover imagine remote shape art doe originally open item
Topic 33	kind	kind break dan trouble age park soul douglas eye guide
Topic 34	post	post won response product thread description familiar doe writer topic
Topic 35	set	set size event character colors reduce default background keyboard parent
Topic 36	isn	isn henry andy listen flight fred remind wear baby doe
Topic 37	long	long orbit quickly doe mix putting foot hole lock lab
Topic 38	call	call peter entire paint frame style table doe solid att
Topic 39	bad	bad comment pat riding damage cop truck sick handling finding
Topic 40	show	show inside straight sex johnson pressure sexual male doe complain