Appendix

In this section, we provide the proof of Proposition 1 as well as more details on the algorithm comparison between sparse topical coding (STC) and the probabilistic LDA [2], hyper-parameter selection/estimation, the implementation of regularized LDA models as evaluated in Section 4, the comparison with non-negative matrix factorization [6] which uses similar Poisson distributions to model word counts, and finally the convergence curves of the coordinate descent algorithms for both STC and MedSTC.

A.1. Proof of Proposition 1

Proof: We consider two cases. First, if $x_0 \geq 0$, then by definition, x_0 is the solution of P_0 and $x^\star = x_0 = \max(0, x_0)$. Second, $x_0 < 0$. Let's assume $x^\star \neq \max(0, x_0)$ and define $\alpha = \frac{x^\star}{x^\star - x_0}$. Then, we have $x^\star > 0$, $0 < \alpha < 1$ and $0 = \alpha x_0 + (1-\alpha)x^\star$. Again, by definition, we have $h(x_0) < h(0)$ and $h(x^\star) < h(0)$. Using these inequalities, we get $\alpha h(x_0) + (1-\alpha)h(x^\star) < h(0) = h(\alpha x_0 + (1-\alpha)x^\star)$. This contradicts the convexity of $h(x)$. Therefore, we have $x^\star = \max(0, x_0)$. □

A.2. Algorithm Comparison

Figure 1 outlines the structure of the coordinate descent procedure of STC and the variational EM algorithm of LDA. We can see that they have very similar structures.

A.3. Hyper-parameter Estimation

For LDA, the hyper-parameter (i.e., the Dirichlet prior parameter α) can be automatically learned using a gradient descent method. For STC, currently we use a generic grid search based on cross-validation to select the hyper-parameters (λ, γ, ρ). In our experiments, we restrict our grid search by setting γ as a function of λ.

A.4. Regularized LDA using an Entropic Regularizer

In this section, we briefly present the regularized LDA using an entropic regularizer.

By assuming a Dirichlet prior over the topic mixing proportion θ, LDA defines a joint distribution $p(\theta, \{z_m\}_{m=1}^M, \mathbf{w}|\alpha, \beta)$ for a document, which is factorized as

$$p(\theta, \{z_m\}_{m=1}^M, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{m=1}^M p(z_m|\theta) p(\vec{w}_m|z_m, \beta),$$

where both the topic assignment model $p(z_m|\theta)$ and the word generating model $p(\vec{w}_m|z_m, \beta)$ are normalized multinomial *distributions* and α are Dirichlet parameters.

Given a document $\mathbf{w}$, the inference is to compute the posterior distribution $p(\theta, \{z_m\}|\mathbf{w}, \beta)$ [1]. However, the inference is intractable in LDA. Therefore, approximate inference algorithms including variational and MCMC [4] methods have been popularly used to perform the inference task. We concentrate on the variational techniques, which can naturally incorporate regularization, as detailed below.

The basic idea of variational methods is to introduce a variational distribution $q(\theta, \{z_m\}_{m=1}^M|\gamma, \phi)$ to approximate the posterior distribution $p(\theta, \{z_m\}|\mathbf{w}, \beta)$, where γ and ϕ are variational parameters. We follow the mean field method [2] and define

$$q(\theta, \{z_m\}|\gamma, \phi) = q(\theta|\gamma) \prod_{m=1}^M q(z_m|\phi_m),$$

where $q(\theta|\gamma)$ is a Drichlet distribution and $q(z_m|\phi_m)$ is multinomial. Then, the variational inference is to find the optimal parameters (γ^*, ϕ^*) that solve the Kullback-Leibler (KL) divergence minimization problem

$$\min_{\gamma,\phi} KL(q(\theta, \{z_m\}_{m=1}^M|\gamma, \phi) \| p(\theta, \{z_m\}_{m=1}^M|\mathbf{w}, \beta). \quad (1)$$

As shown in [2], a coordinate descent method solving the problem (1) gives the following update equations

$$\phi_{mk} \propto \exp\left\{\mathbb{E}_q[\log(\theta_k)|\gamma] + \log \beta_{kw_m}\right\} \quad (2)$$

$$\gamma_k = \alpha_k + \sum_{m=1}^M \phi_{mk}, \quad (3)$$

where we have used w_m to denote the term id appearing at position m, that is, $\vec{w}_{mw_m} = 1$.

Then, the regularized LDA is to solve the regularized KL-divergence minimization problem [2]

$$\min_{\gamma,\phi} KL(q(\theta, \{z_m\}_{m=1}^M|\gamma, \phi) \| p(\theta, \{z_m\}_{m=1}^M|\mathbf{w}, \beta) + \lambda H(q(\theta, \{z_m\}_{m=1}^M|\gamma, \phi)), \quad (4)$$

[1] The collapsed methods [8] integrates out the document-wise mixing proportion θ by exploring the conjugateness. We choose to directly infer the mixing proportion, which is useful for many applications, such as document classification. But in principle, we can also perform the collapsed variational inference [8] with an entropic regularizer, which could potentially be more efficient than the standard variational method [1].

[2] Note that the entropic regularizer is put on the variational distributions instead of the original model posterior distribution. However, by minimizing the KL-divergence, we can expect to project the original model distribution to a space with desired properties (e.g., sparsity) and therefore introduce appropriate regularization to the original model [3].

Coordinate Descent Alg. of STC

Input: corpus $\mathcal{D} = \{\mathbf{w}_d\}_{d=1}^D$, regularization constants (λ, γ, ρ), topic number K.
Output: distributional topics β, sparse codes θ and $\mathbf{s}$
repeat
  /**** Hierarchical Sparse Coding ****/
  for $d = 1$ **to** D **do**
    for each word $n \in I_d$ **do**
      Update word code $\mathbf{s}_{dn}$.
    end for
    Update document code θ_d.
  end for
  /**** Dictionary Learning ****/
  Update the distributional topics β.
until convergence

Variational EM Alg. of LDA

Input: corpus $\mathcal{D}' = \{\mathbf{w}_d\}_{d=1}^D$, Dirichlet prior parameter α, topic number K.
Output: distributional topics β, approximate posterior distributions $q(\theta|\mathbf{w})$ and $q(z_m|\mathbf{w})$.
repeat
  /**** E-Step ****/
  for $d = 1$ **to** D **do**
    for each position $m = 1$ **to** M_d **do**
      Update topic assignment distribution $q(z_m|\mathbf{w}_d)$.
    end for
    Update the distribution $q(\theta|\mathbf{w}_d)$.
  end for
  /**** M-Step ****/
  Update the distributional topics β.
until convergence

Figure 1: Algorithm comparison of STC and probabilistic LDA, where the left part is the coordinate descent algorithm of STC and the right part is the variational EM algorithm of LDA.

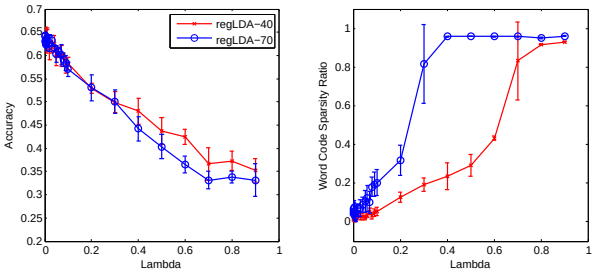


Figure 2: (L) classification accuracy and (R) sparsity ratio of word codes of the regularized LDA with respect to the regularization constant λ, whose values are $0.001, 0.002, \cdots, 0.01, 0.02, \cdots, 0.1, 0.2, \cdots, 0.9$. We show two LDA models with 40 topics and 70 topics.

where $H(p(x)) = -\int_{x\in\mathcal{X}} p(x)\log p(x)$ is the entropy of a distribution $p(x)$ and λ is a non-negative regularization constant. Since a sparser distribution has a smaller entropy than a denser distribution (e.g., uniform distribution which has the largest entropy), minimizing an entropic regularizer will drive the variational distributions to be sparse.

For problem (4), we can easily derive the update equations for γ and ϕ because of the differentiability of the entropic regularizer. Specifically, we have the regularized update equations as follows

$$\phi_{mk} \propto \exp\left\{\frac{\mathbb{E}_q[\log(\theta_k)|\gamma] + \log\beta_{kw_m}}{1-\lambda}\right\} \quad (5)$$

$$\gamma_k = \frac{\alpha_k + \sum_{m=1}^M \phi_{mk} - \lambda}{1-\lambda}. \quad (6)$$

Therefore, to make the variational distribution valid, we need to constrain that $\lambda \in [0, 1)$. Figure 2 shows the classification accuracy and sparsity ratio of the regularized LDA on the 20 Newsgroup dataset when the topic number is set at 40 and 70, respectively. For these models, we automatically estimate the optimal Dirichlet parameter α using the Newton-Raphson method [2]. We can see that when using a strong regularizer (i.e., λ is close to 1), the sparsity ratio will be increased; however, the classification accuracy is dramatically decreased (please also see the performance of regLDA$^-$ in the Figure 5 of the main paper). When λ is small, the classification accuracy is improved a bit as shown in Figure 5 (regLDA$^+$) of the main paper, but the sparsity ratio is still very small.

A.5. Comparison with Non-negative Matrix Factorization

As we have discussed in Section 1, STC is related to the non-negative matrix factorization [6]. Let $\mathbf{X}$ denote the observed $N \times D$ word count matrix, where rows represent terms in a dictionary and columns represent documents. Then, non-negative matrix factorization (NMF) is to find non-negative matrices $U \in \mathbb{R}^{N\times K}$ and $V \in \mathbb{R}^{K\times D}$ such that $\mathbf{X} \approx UV$, where K is the rank which is usually much smaller than N. Each column of the matrix U represents a basis and each column of V is the non-negative coefficient vector that is used to reconstruct *all* the observed word counts in a document. In [6], a similar log-Poisson loss is used to estimate matrices U and V.

STC is significantly different from NMF, analogous to the difference between latent Dirichlet allocation (LDA) and mixture of unigrams [2]. First, NMF uses one document-specific coefficient vector to reconstruct all the observed word counts in the same document. This assumption is often too limiting to effectively model a large collection of documents. In contrast, STC allows different words in one document to exhibit different sparsity patterns via using different word codes. Second, for each document, NMF (as well as the sparse coding method [7]) aims to reconstruct a vector with all the words in a vocabulary, while STC only reconstructs the words with non-zero counts. Us-

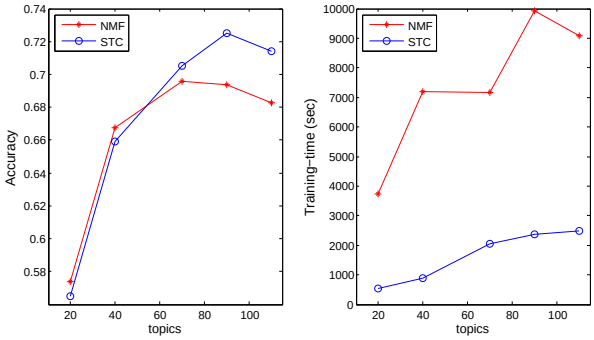


Figure 3: (L) classification accuracy and (R) training time of NMF and STC when using different number of topics.

ing the sparse representation could make STC more efficient and scalable to a large vocabulary.

Empirically, the sparsity ratio of the "word codes" (i.e., the document-specific coefficient vector, which is the same for all the words in a document; we can call them "document codes" as well.) is much smaller (around 0.005 for different numbers of topics ranging from 10 to 110) than the word code sparsity ratio (or the document code sparsity ratio) of STC or even gaussSTC which uses ℓ_2-norm regularizers. Therefore, NMF is limiting in using one document-specific coefficient vector to reconstruct all the word counts in that document and cannot identify the sparse topical meanings of each individual words. Although using a sparsity-inducing constraint [5] can improve the sparseness of the coefficient vector in NMF, it still cannot identify the sparse topical meanings of each individual words because of the intrinsic limitation. Moreover, as shown in Figure 3, NMF performs worse than STC on classification accuracy when the topic number is large (e.g., larger than 60) and the standard multiplicative algorithm [6] is much more expensive than our coordinate descent algorithm for training STC. In this experiment, we actually only consider the words with non-zero counts to make NMF scalable to the vocabulary for 20 Newsgroup data, which contains more than 60,000 words.

A.6. Comparison between Using ℓ_1 and ℓ_2-norm on θ

Figure 4 shows the word-code sparsity ratio of STC and MedSTC when using ℓ_1-norm or ℓ_2-norm on the document code θ. We can see that using ℓ_2-norm leads to a sparser word code in STC than that obtained by using an ℓ_1-norm. For MedSTC, the sparsity ratios of using two different norms are comparable.

Figure 5 shows the change of objective values and log-Poisson loss (i.e., negative log-Poisson likelihood) of STC during training. We compare the STC using ℓ_1-

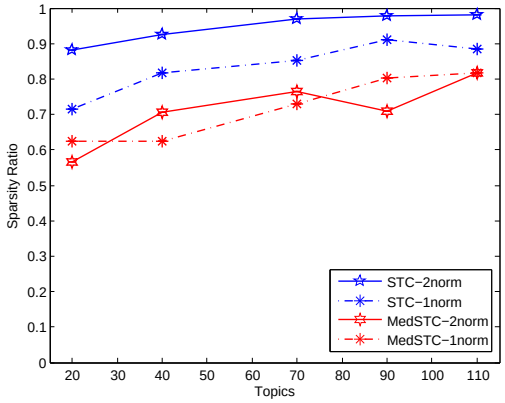


Figure 4: Word code sparsity ratio of STC and MedSTC when using ℓ_2-norm or ℓ_1-norm on the document code θ.

norm or ℓ_2-norm on the document code θ. We can see that the coordinate descent algorithm (for both types of regularizers) is very stable and converges very fast. In most cases, about 100 iterations are good enough. We set the stopping criterion as the absolute relative change of objective function is less than $1e^{-5}$, and the maximum iteration number is 100.

Similarly, the coordinate descent algorithm for MedSTC using ℓ_1-norm regularizer on θ is also very stable and converges fast, as shown in Figure 6 (a), where we set $\ell = 360$ and $\gamma = 10\lambda$ when doing the constrained parameter searching. For the ℓ_2-norm regularizer on θ, the objective value shows some disturbance, mainly due to the non-smoothness of SVM hinge loss and the larger ℓ used. But the log-Poisson loss curves under both cases are decreasing stably. Comparing the log-Poisson loss between STC and MedSTC, we can see that using supervised side information (i.e., class labels) can improve the fitness of the model.

References

[1] A. Asuncion, M. Welling, P. Smyth, and Y.W. Teh. On smoothing and inference for topic models. In *UAI*, 2009.

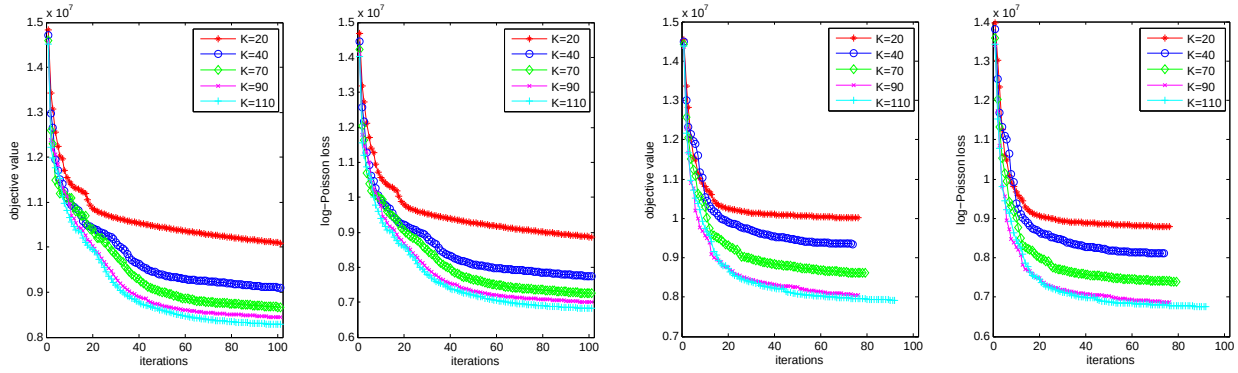
[2] D. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *JMLR*, (3):993–1022, 2003.

[3] K. Ganchev, J. Graa, J. Gillenwater, and B. Taskar. Posterior regularization for structured latent variable models. *JMLR*, (11):2001–2094, 2010.

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

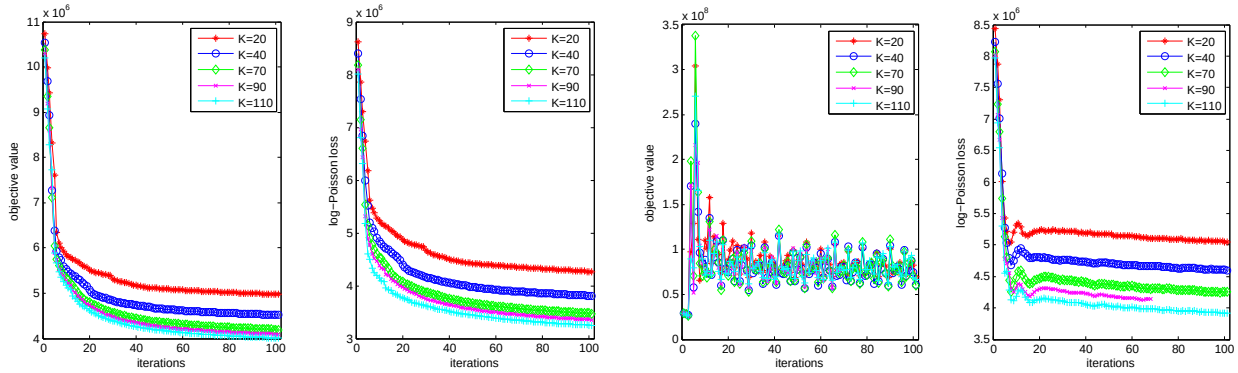
[5] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *JMLR*, (5):1457–1469, 2004.

[6] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788 – 791, 1999.



(a) (b)

Figure 5: (**a**) objective function value and log-Poisson loss of STC using ℓ_1-norm on document code θ; and (**b**) objective function value and log-Poisson loss of STC using ℓ_2-norm on document code θ, when using different number of topics.



(a) (b)

Figure 6: (**a**) objective function value and log-Poisson loss of MedSTC using ℓ_1-norm on document code θ; and (**b**) objective function value and log-Poisson loss of MedSTC using ℓ_2-norm on document code θ, when using different number of topics.

[7] H. Lee, R. Raina, A. Teichman, and A.Y. Ng. Exponential family sparse coding with applications to self-taught learning. In *IJCAI*, 2009.

[8] Y.W. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *NIPS*, 2006.