Concave regularizations and MAP priors for sparse topic models

Johan H. Ugander
Center for Applied Mathematics
Cornell University
jhu5@cornell.edu
[Report for CS6780: Advanced Machine Learning, Fall 2010]

1 Introduction

Across all sectors of the modern information economy, large unstructured repositories of data are being aggregated at an ever-increasing rate. This move towards 'big data' has created an enormous demand for techniques to efficiently extract structure from such data sets. Specific contexts for this demand include natural language models for organizing text corpuses, image feature extraction models for navigating large photo datasets, and community detection in social networks for optimizing content delivery. Models of such structure are broadly called *topic models* or *latent variable mixture models*, aiming to identify maximally informative latent topics common to different elements of the unstructured dataset.

In this work our primary context for topic modeling will be natural language processing. In this context we will regard documents d belonging to a fixed collection D ($|D| = m$) via the simplifying assumption that each document is merely an unordered collection of words w drawn with replacement from a fixed vocabulary V ($|V| = n$). From this assumption we obtain a $n \times m$ dimensional data matrix X of word counts, where element X_{ij} is the number of times word w_i appeared in document d_j.

Given such a data matrix, the goal of natural language topic models is to find a decomposed matrix approximation of X such that each column of X corresponding to a document is interpretable as a linear combination $\sum_{k=1}^{K} w_k h_{kj}$, where the vectors w_k are the vocabularies of K topics, and the elements h_{kj} are the topic weights associating document d_j with each topic. Generally, we seek an $n \times K$ vocabulary matrix W and a $K \times m$ document matrix H such that X is well-approximated by WH.

Ordinarily it is assumed that $K \ll m$, namely that the number of topics is much less than the number of documents. We call this the *low-rank assumption*. In this work we explore an additional assumption with notably different consequences, namely that individual document only incorporate a small subset of the k topics, and as such each document is assumed to arise as a mixture of only $L \ll K$ of the topics. This is equivalently an assumption about the sparsity of the document matrix H, and so we call this the *sparsity assumption*. Topic models are often celebrated for their ability to naturally infer sparse topic weights for documents without any assumption necessary, but in this work we perform a formal analysis of how this assumption can be harnessed explicitly to improve the performance of topic models.

Common algorithms for topic modeling discussed in this work include Non-negative matrix factorization (NMF, [12]), Probabilistic latent semantic indexing (PLSI, [11]), and Latent dirichlet allocation (LDA, [4]). The three algorithms are in fact very closely related: NMF is an unassuming optimization problem, PLSI is a nearly equivalent probabilistic re-formulation of NMF [9], and LDA is fully Bayesian generative extension of PLSI.

The sparsity assumption we make is fundamentally a frequentist assumption, since sparsity is ultimately a statement about a single 'maximum likelihood' or 'maximum a posteriori' point-estimated topic model produced as the output of some optimization scheme over the class of models. Meanwhile fully Bayesian approaches such as LDA output posterior distributions across the model space. As a result, our discussion of LDA is limited to occasional commentary, and the main focus of our work is the inference schemes for ML and MAP estimated NMF and PLSI models.

The paper is organized as follows. In section 2, we present the topic models behind traditional NMF and PLSI, derived as maximum likelihood estimation problems for probabilistic models, and establish the connections between NMF and PLSI previously observed in the literature. In section 3 we extend NMF and PLSI towards the MAP framework with the goal of incorporating a sparsity assumption. Here we identify equivalencies between regularizations of the ML optimization problem and MAP prior distributions. In section 4, we present an application of our MAP/regularization techniques for text mining web logs.

2 ML Point Estimation

This section reviews point estimation for NMF and PLSI topic models, based on previously established results from the research literature.

2.1 NMF as Maximum Likelihood

As an optimization problem, non-negative matrix factorization seeks $W \in \mathbb{R}_+^{n\times k}$, $H \in \mathbb{R}_+^{k\times m}$ such that some distance between X and WH is minimized. While many distances have been considered in the literature [8, 12], we will principally concern ourselves with minimizing the Kullback-Leibler (KL) divergence:

$$D(X||WH) = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}\log\frac{X_{ij}}{(WH)_{ij}} - X_{ij} + (WH)_{ij}. \tag{1}$$

Note that $\sum_{ij} X_{ij}$ is constant. With this, the problem of minimizing the KL divergence becomes:

$$\min_{W,H}\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}\log\frac{X_{ij}}{(WH)_{ij}} + (WH)_{ij}, \qquad \text{s.t. } W,H \geq 0 \tag{2}$$

This objective function is non-convex, which implies that global maxima will be very hard to find with any certainty.

The specific interest for NMF measured by KL divergence, henceforth referred to simply as NMF, is due to it's exact agreement with the maximum likelihood inference problem of estimating the parameters of a Poisson random matrix [13]. To see this, let X_t be an observation of a k-topic random matrix $\Pi_t \sim \text{Pois}(tWH)$, where each element $(\Pi_t)_{ij} \sim \text{Pois}(t\sum_{k=1^K} w_{ik}h_{kj})$. Here W, H are our low-rank rectangular matrixes as above, where $w_{ik} \geq 0$ is the intensities for the word i in topic k and $h_{kj} \geq 0$ designates the 'amount' of topic k in document j. Let $X = t^{-1}X_t$ be a time-normalized version of the observation.

Given a data matrix X originating from this model, we can derive the Maximum Likelihood estimate of the intensity parameter matrixes $W, H \geq 0$. The log-likelihood function for the model is given by:

$$\ell(W,H;X) = \sum_{ij}\log\left(\frac{(WH)_{ij}^{X_{ij}}}{X_{ij}!}\exp\{-(WH)_{ij}\}\right) = -\sum_{i,j} X_{ij}\log\frac{X_{ij}}{(WH)_{ij}} - (WH)_{ij}, \tag{3}$$

where the constant term has been dropped. We see that maximizing the log-likelihood is equivalent to the NMF minimization problem in (2). The preferred algorithm for solving this problem within the NMF literature is a multiplicative update algorithm derived by Lee and Seung [12] via a manipulation of gradient descent, derived here for W_{ik} only:

$$W_{ik} \leftarrow W_{ik} + \delta_{ik}\Big(\sum_j \frac{X_{ij}H_{kj}}{(WH)_{ij}} - \sum_j H_{kj}\Big).\ \forall i,k, \tag{4}$$

With a clever choice of step-size $\delta_{ik} = W_{ik} / \sum_j H_{kj}$ and a corresponding choice of step-size for updating all H_{kj}, we obtain a set of multiplicative update rules, which, given non-negative initial conditions, are guaranteed to remain non-negative:

$$W_{ik} \leftarrow W_{ik} \frac{\sum_j H_{kj} \frac{X_{ij}}{(WH)_{ij}}}{\sum_j H_{kj}}, \ \forall i,k \qquad H_{kj} \leftarrow H_{kj} \frac{\sum_i W_{ik} \frac{X_{ij}}{(WH)_{ij}}}{\sum_i W_{ik}}. \ \forall j,k \tag{5}$$

The algorithm is guaranteed to converge to a local optima [12].

2.2 Constrained NMF for identifiability

Given a solution (W, H) to (2), any transformation of the form $(WB, B^{-1}H)$ will obtain the same value of the objective function. During implementation, it is therefore convenient to fix the ℓ_1-norm for the columns or rows of one of the matrixes, typically the columns of W. This is effectively done by normalizing the column sums, projecting the column vectors of W onto the simplex. The constrained algorithm is then:

$$W'_{ik} \leftarrow W_{ik} \frac{\sum_j H_{kj} \frac{X_{ij}}{(WH)_{ij}}}{\sum_j H_{kj}}, \ \forall i,k \qquad H_{kj} \leftarrow H_{kj} \frac{\sum_i W_{ik} X_{ij}}{(WH)_{ij}}, \ \forall j,k \tag{6}$$

$$W_{ik} \leftarrow W'_{ik} / \sum_i W'_{ik} \ \forall k. \tag{7}$$

Equivalently, this normalization is what results from considering the constrained optimization problem via the method of Lagrangian multipliers (not shown). This modification preserves both the non-negativity and the convergence properties of the original algorithm [9].

2.3 Maximum Likelihood PLSI

Following [9, 10], we briefly review the equivalence between NMF and PLSI. Let $Y = X/(\sum_{ij} X_{ij})$ be a normalization of the data matrix. Probabilistic latent semantic indexing approaches Y as an empirical observation of the joint distribution $P(i, j)$ over (word, document) pairs, and assumes that both word and document distributions are latent variable mixtures of conditional independent multinomial distributions. PLSI seeks to identify a decomposition:

$$P(w_i, d_j) = \sum_{k=1}^{K} P(w_i, d_j, z_k) = \sum_{k=1}^{K} P(w_i|z_k)P(z_k)P(d_j|z_k), \tag{8}$$

where z_k are the latent variables corresponding to the K topics. Given Y as an observation of $P(w_i, d_j)$ and regarding $P(w_i|z_k), P(d_j|z_k)$, and $P(z_k)$ as parameters of the model, the log-likelihood function for these parameters can be manipulated as:

$$\ell = \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} \log P(w_i, d_j) = -\sum_{ij} Y_{ij} \log \frac{Y_{ij}}{P(w_i, d_j)} - P(w_i, d_j), \tag{9}$$

where we use the fact that $\sum_{ij} Y_{ij} \log Y_{ij}$ and $\sum_{ij} P(w_i, d_j) = 1$ are constants. We recognize the log-likelihood maximization problem for Y as corresponding to the NMF minimization for X in (2). Since X and Y differ only by a multiplicative constant, the optimization problems are the same. This leads to the following proposition.

Proposition 1. *Probabilistic Latent Semantic Indexing for inferring an approximate decomposition $P(w_i, d_j) = \sum_k P(w_i|z_k)P(z_k)P(d_j|z_k)$ is equivalent to Non-negative Matrix Factorization for inferring an approximate decomposition $X = WH$ for a normalized matrix X.*

Proof. Setting $W_{ik} = P(w_i|z_k)$, we see that W is exactly the simplex-normalized vocabulary matrix we considered when solving NMF. Setting $H_{kj} = P(z_k)P(d_j|z_k)$, it can be shown (see [9]) that since Y is normalized and W is always column-stochastic, H returned from the NMF multiplicative update algorithm will also be normalized. As such it can be decomposed into a row-stochastic matrix $P(d_j|z_k)$ and a $k \times k$ normalized diagonal matrix $P(z_k)$.

Since the NMF and PLSI objective functions are the same, and a solution to the NMF problem yields a solution to the PLSI problem (it is straight-forward to see that the reverse is also true), they are the same problem. □

While PLSI can thus be solved by the multiplicative update rules of NMF, PLSI is more typically solved by the EM algorithm commonly used for probabilistic mixture models. The EM algorithm has the same stationary conditions but in fact follows a different trajectory through the fitness landscape [10], an observation that has notably led to the development of hybrid algorithms [9].

3 Priors and regularization

This section offers novel contributions to the NMF and PLSI literature by considering MAP and regularized formulations of the traditional ML point estimation problem, specifically adapted for inferring sparse topic models.

3.1 On regularization for sparsity

In order to formally impose sparsity constraints on our document matrix H in our NMF problem, or similarly impose sparsity on the document-topic distribution $P(\mathbf{d}|\mathbf{z})$ from PLSI, we would need to solve the optimization problem in (2) subject to a constraint on the ℓ_0 norm of H. Equivalently (by constructing an unconstrained Lagrangian), we could penalize/regularize the objective function in (2) by adding a regularization term $\lambda||H||_0$, for some $\lambda > 0$. The problem with ℓ_0 regularization is that the ℓ_0 norm is combinatorial in nature, and there are 2^{mK} different sparsity patterns for the H matrix that would need to be considered.

This difficulty is commonly circumvented by considering the ℓ_1 norm as an approximation of the ℓ_0 norm, and imposing an ℓ_1 regularizations of the model parameters, as was pioneered by the LASSO algorithm for regression shrinkage [15]. The strength of ℓ_1 regularization is normally that it can be added to convex objective functions (such as least squares for regression) to obtain a quadratic program that can be solved efficiently. Recall that the KL-NMF objective function is not convex, and so in our context we do not necessarily gain anything form this point.

The intuition behind the sparsity resulting from ℓ_1 regularization is that the gradient of the penalty term is discontinuous at zero, and persistently penalizes even small values towards zero, which preferentially selects for sparse optima. For comparison, consider ℓ_2 regularization (as found in ridge regression [5]), where the penalty gradient has a vanishing effect on the overall gradient movement as parameter values approach zero.

In the original derivation of the LASSO algorithm, Tibshirani further noted that ℓ_1 regularization corresponds to performing a MAP estimation of the regression parameters given a double-exponential prior on the parameters. Contrasting this to the zero-mean Normal prior corresponding to ℓ_2-regularized ridge regression, the density of the double-exponential prior is more concentrated at zero *and* in its tails, leading to a greater concentration around a few large values [15].

At this point it is important to note that the parameters of PLSI, namely the distributions we are estimating, $P(\mathbf{z})$, $P(\mathbf{w}|z_k)$ and $P(\mathbf{d}|z_k)$ for $k = 1, \ldots, K$, are all restricted to taking values on some simplex, i.e. each such distribution has a fixed ℓ_1 norm, with the total ℓ_1 norm of the document-topic distribution $P(\mathbf{d}|\mathbf{z})$ fixed at K. This makes ℓ_1 regularization inapplicable for PLSI. In order to induce sparsity on our PLSI inference problem, we are forced to consider strictly concave regularization functions.

First we will consider correspondences between MAP priors and regularizations for NMF, where the document matrix H is not restricted in its ℓ_1 norm. In doing so, we will see that a specific concave regularization (log-sum) has a favorable MAP interpretation. We then develop this regularization further for the case of PLSI, proposing a novel algorithm for inducing sparsity on PLSI point estimates.

3.2 MAP NMF: Correspondences between priors and regularizations

For MAP estimation, we consider prior distributions on H, characterized by a density f_H with hyperparameters θ. We aim to find the $W, H \geq 0$ which maximize the log-likelihood function:

$$\ell_{MAP} = \log\Big(\prod_{ij} Pr(X_{ij}|W,H) \times \prod_{jk} f_H(H_{kj}|\theta)\Big) = \ell_{ML} + \sum_{jk} \log f_H(H_{kj}|\theta_H) \qquad (10)$$

We identify $g(H_{kj};\theta) = -\sum_{jk} \log f_H(H_{kj}|\theta)$ as the MAP-equivalent regularization term, and from this we can make the general observation that MAP estimation corresponds to the regularized optimization problem

$$\min_{W,H} D(X||W,H) + g(H_{kj};\theta) \qquad \text{s.t. } W,H \geq 0, \sum_i W_{ik} = 1, \forall k. \tag{11}$$

This is in fact not a statement about NMF, but about the relationship between MAP estimation and regularization in general. We are specifically interested in prior distributions over the non-negative real numbers $\mathbb{R}_+$ with a high likelihood for taking a value of zero. In light of this, we present the following correspondences.

Proposition 2. *MAP estimation with the prior $H \sim Exp(\theta)$ i.i.d. is equivalent to ℓ_1-regularized optimization, where $g(H_{kj};\theta) = \theta||H||_1$.*

Proposition 3. *MAP estimation with the prior $f_H(x|\alpha) = (1+x)^{-\alpha-1}\mathbf{1}_{[x\geq 0]}$, e.g. $H \sim Pareto_{[0,\infty)}(\alpha)$ i.i.d., is equivalent to regularized optimization where $g(H_{kj};\theta) = (1+\alpha)\sum_{kj}\log(1+H_{kj})$.*

The correspondence between the exponential prior and ℓ_1 regularization is merely the non-negative version of the correspondence observed for the LASSO algorithm, and in fact this correspondence for NMF has been alluded to before in an earlier examination of MAP NMF not focussed on sparsity [7]. More interestingly, we notice that placing a heavy-tailed Pareto prior on H corresponds to a concave log-sum regularization of the elements. With our goal of sparsity in mind, this penalty function is in fact a better approximation than ℓ_1 to ℓ_0.

Concave regularizations are of little interest for convex optimization problems, since they breaks the common efficient solution methods using convex programming, but given that our problem is not convex, it is fully deserving of our consideration. Log-sum regularization is particularly deserving of our consideration when we consider the recent work by Candès, Wakin, and Boyd on weighted ℓ_1 minimization for compressed sensing [6].

The work by Candès et al. explores improvements upon the sparsity properties of ℓ_1 regularization, proposing an algorithm for iteratively solving a sequence of adaptively re-weighted ℓ_1 minimization problems in order to 'more democratically penalize non-zero coefficients.' The authors show that their iteratively re-weighted minimization outperforms ordinary ℓ_1 minimization for reconstructing sparse signals in a variety of compressed sensing problems. Importantly, they show how their iteratively reweighted ℓ_1 algorithm is in fact a majorization-minimization (MM) algorithm for solving the non-convex optimization problem

$$\min \sum_{i=1}^{n} \log(|x_i| + \epsilon) \qquad \text{s.t. } \Phi x = y, \tag{12}$$

where Φ is a matrix specified by the compressed sensing problem and $\epsilon > 0$ is a smoothing parameter of their algorithm governing the severity of the regularization. Given the success of their algorithm for compressed sensing, log-sum regularization is a promising approach to inducing sparsity, and to the best of our knowledge the correspondence between Pareto priors and $\epsilon = 1$ in the Candès et al. algorithm is novel.

Moreover, this connection suggests a concave regularization algorithm for inducing sparsity on the simplex-constrained PLSI parameters, which we consider in Section 3.4.

3.3 A multiplicative MAP NMF algorithm

Here we derive an intuitive sufficient condition on the prior probability distribution f_H such that the multiplicative update algorithm will remain non-negative.

Lemma 1. *If the prior densities f_H is non-increasing, i.e. $f'_H(x|\theta) \leq 0$, $\forall x \geq 0$, then the non-negative multiplicative update algorithm for MAP is given by*

$$W'_{ik} \leftarrow W_{ik}\frac{\sum_j H_{kj}\frac{X_{ij}}{(WH)_{ij}}}{\sum_j H_{kj}}, \ \forall i,k, \qquad H_{kj} \leftarrow H_{kj}\frac{\sum_i W_{ik}\frac{X_{ij}}{(WH)_{ij}}}{1+g'(H_{jk};\theta)}, \ \forall j,k, \tag{13}$$

$$W_{ik} \leftarrow W'_{ik}/\textstyle\sum_i W'_{ik} \ \forall k$$

where $g(H_{kj};\theta) = -\sum_{jk}\log f_H(H_{kj}|\theta)$.

Proof. The multiplicative update algorithm follows from constructing the induced gradient descent algorithm, following (4), and setting the step-size for H_{kj} to $\delta_{ik} = H_{kj}/(1 + g'(H_{jk}; \theta))$.

It is clear that we require the denominator of the multiplicative update factor for all H_{kj} to be non-negative. Recall that the density is non-negative, $f_H(H_{kj}|\theta) \geq 0$. From the following manipulation

$$1 + g'(H_{jk}; \theta) = 1 - \frac{\partial}{\partial H_{kj}} \log\left(f_H(H_{kj}|\theta)\right) \geq 0, \forall j, k \quad \Leftrightarrow \quad \frac{f'_H(H_{kj}|\theta)}{f_H(H_{kj}|\theta)} \leq 1, \forall j, k, \tag{14}$$

we see that $f'_H(H_{kj}|\theta) \leq 0$ is clearly sufficient. □

In the previous section we noted that we were specifically interested in priors with a high likelihood for taking a value of zero. Fortunately, this is precisely what $f'_H(H_{kj}|\theta) \leq 0$ says, and from the above lemma we see that both the Exponential and Pareto priors yield non-negative MAP update algorithms. Note that we have not established the convergence of this algorithm at this time.

3.4 MAP PLSI and the Bounded pseudo-Dirichlet distrubtion

In Section 2.3 we showed that for maximum likelihood estimation, PLSI is equivalent to a normalized version of NMF. For the purposes of maximum likelihood inference, it is sound to assume that the normalized document matrix H can be decomposed as $H = P(\mathbf{z})P(\mathbf{d}|\mathbf{z})$. For MAP inference, it is however important to note that the normalization and decomposition assumptions greatly restrict the possible prior distributions one can tractably approach. For instance, the Pareto i.i.d. prior for the elements of H that is feasible for NMF is not tractable for PLSI: the elements of each distribution $P(\mathbf{d}|z_k)$ have a non-trivial joint structure on a $(m-1)$-dimensional simplex with no clear Pareto-i.i.d.-analog.

For MAP PLSI we are restricted to considering prior distributions on the simplex, with the canonical variety being the class of Dirichlet distributions Dir($\vec{\alpha}$). As a well-balanced prior distribution, we focus our interests on the special case of the symmetric Dirichlet distribution Dir($\alpha\mathbf{1}$), where α is the scalar *concentration parameter* which corresponds to how concentrated we think the distribution is on the simplex. A prior with a small concentration parameter ($\alpha \ll 1$) assumes that the distribution is concentrated on only a few elements, while a large concentration parameter ($\alpha \gg 1$) assumes a relatively even distribution across the elements. As such, the concentration parameter is exactly a sparsity parameter.

Recall that the density function of the symmetric Dirichlet distribution is given by

$$f_D(x_1, \ldots, x_N; \alpha\mathbf{1}) = \frac{1}{\mathbf{B}(\alpha\mathbf{1})} \prod_{i=1}^{N} x_i^{\alpha-1}, \qquad \mathbf{x} \in \Delta^{N-1}, \tag{15}$$

where $\mathbf{B}(\cdot)$ is the multinomial Beta function and Δ^{N-1} is the $N-1$ dimensional simplex. From the above presentation, the Dirichlet distribution comes across as the perfect prior distribution for inducing sparsity on the simplex. Indeed, Latent dirichlet allocation (LDA) takes precisely this approach [4].

The problem is that the Dirichlet density function is unbounded precisely for sparse concentration parameters $\alpha < 1$, which makes point estimation with a concentrated Dirichlet prior an ill-defined optimization problem. Previous work on MAP PLSI has dodged this problem by focusing on Dirichlet priors that introduce smoothing ($\alpha > 1$) rather than sparsity [1], in which case the density is indeed bounded. Also, the fully Bayesian framework behind LDA is also capable of overcoming the unbounded density function, since LDA seeks a posterior distribution rather than a point-estimate, but as we've mentioned before, the downside of this is that LDA can not be trained to find true sparsity.

We address the Dirichlet distribution's unbounded density by proposing a distribution that is a modification specifically intended for use in point-estimating sparse parameter vectors on the simplex. We call this distribution the *bounded pseudo-Dirichlet distribution*.

Definition 1. *The bounded pseudo-Dirichlet distribution has the probability density function*

$$f_D(x_1, \ldots, x_N; \alpha\mathbf{1}) = C(\epsilon) \prod_{i=1}^{N} (x_i + \epsilon)^{-1} \qquad \mathbf{x} \in \Delta^{N-1}, \tag{16}$$

where $\epsilon > 0$ is the regularization parameter, $C(\epsilon)$ is a normalization constant that depends on ϵ, and Δ^{N-1} is the $N-1$ dimensional simplex.

Notice the close similarity to the Dirichlet distribution in (15), where we have simply set $\alpha = 1$ and used ϵ to smooth the unbounded singularities along the simplex boundary. Setting $\alpha = 1$ is not strictly necessary, but reduces the distribution to a one parameter family. More significantly, the bounded pseudo-Dirichlet prior corresponds exactly the ϵ-smoothed log-sum penalty which Candès et al. demonstrated as a concave improvement over ℓ_1 optimization for compressed sensing [6]. We formalize this as a proposition.

Proposition 4. *MAP point estimation with a Bounded pseudo-Dirichlet prior distribution BpD(ϵ) on the simplex is equivalent to regularized optimization where $g(H_{kj}; \epsilon) = \sum_{kj} \log(H_{kj} + \epsilon)$.*

The shape of the bounded pseudo-Dirichlet distribution is shown in Figure 1, along with corresponding log-sum regularizations. The BpD(ϵ) prior lacks many of the formal properties of the Dirichlet distribution. Notably, unlike the Dirichlet distribution it is not a conjugate prior to the multinomial distribution. But for the purposes of regularized point-estimation of parameters on a simplex, it is a very clean parameterization of prior knowledge for sparsity, and unlike the Dirichlet distribution, the fact that it is bounded leads to a well-posed MAP optimization problem.

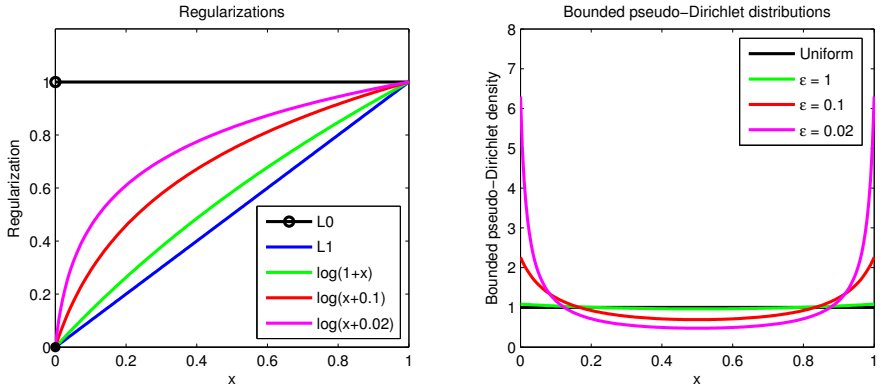

Figure 1: Left: the relative concavity of different sum-log regularization functions, which have been translated and rescaled for comparison. Right: The 2-dimensional Bounded pseudo-Dirichlet distribution on $[0, 1]$ (the 1-dimensional simplex), for several values of ϵ.

4 Application: text mining weblogs

We present a very brief application of the theory developed above, in which we infer a PLSI topic model using a Bounded pseudo-Dirichlet MAP-gradient descent algorithm for a medium sized corpus of 2894 blogger.com blogs, a dataset originally analyzed by Schler et al. [14] for the role of gender in language. Unigram frequencies for the blogs were built using the python NLTK toolkit [3]. Inference was run on a document-word matrix consisting of the 2894 blogs and 5000 most common words, as determined by the aggregate frequencies across the entire corpus. The blogs in the corpus all had associated with them user-provided labels, providing one of 28 categories. We focused our analysis on 10 varied but representative topics, while the complete corpus contained over 19,000 blogs.

Ordinary NMF gradient descent descent was compared to Bounded pseudo-Dirichlet MAP gradient descent, where the regularization parameter was set to $\epsilon = 0.02$. A 5-fold cross-validation to determine an optimal ϵ would have been preferred, and an investigation into the nature of ϵ remains as future work. The MAP algorithm was initiated from the solution of the ML NMF algorithm. This follows the observation of Candès et al. [6] that the concave regularization can attenuate local optima, and so starting from the solution to the ML problem ensures that the algorithm has reached a reasonable neighborhood of the solution space. The NMF algorithm was initiated with a uniformly drawn initial condition.

The ground truth topic labels in the data are quite noisy, and really we feel that this is perhaps not the ideal data set to publish this algorithm for, so a quantitative analysis of performance was not pursued. From Figure 2, it is however qualitatively clear that the prior does indeed induce sparsity. The rows of Figure 2 represent the average of the topic distributions for documents with the label found in that row. Notice also that between the ML optima and the MAP optima, the topics have subtly shifted, as if to enable sparser topics. For example, 'republican' shifts into the leading terms of the 'law' basis, allowing religion to find a sparser topic distribution.

5 Conclusion

Many probabilistic inference problems such as PLSI are constrained in their ℓ_1 norm, making traditional ℓ_1 regularization toothless for sparse inference. In this work, we have presented a well-motivated concave regularization for topic models, and described it's corresponding prior distribution, a close but novel relative of the symmetric Dirichlet distribution that we term the Bounded pseudo-Dirichlet. Using this prior for MAP estimation encourages the selection of sparse topic models, which we demonstrate via an application of our MAP algorithm for text mining, where we observe qualitatively impressive results.

References

[1] A. Asuncion, M. Welling, P. Smyth, Y.W. Teh."On smoothing and inference for topic models," *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 27–34, 2009.

[2] M.W. Berry, M. Brown. "Email Surveillance Using Non-negative Matrix Factorization," *Computational & Mathematical Organization Theory*, vol. 11, pp. 249?264, 2005.

[3] S. Bird, E. Klein, E. Loper, *Natural language processing with Python*, O'Reilly Media, 2009.

[4] D.M. Blei, A.Y. Ng, M.I. Jordan. "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] C.M. Bishop, *Pattern recognition and machine learning*, Springer, New York, 2006.

[6] E.J. Candès, M.B. Wakin, S.P. Boyd. "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, pp. 877–905, 2008.

[7] A.T. Cemgil. "Bayesian Inference for Nonnegative Matrix Factorisation Models," *Computational Intelligence and Neuroscience*, vol. 2009, article 785152, 2009.

[8] A. Cichocki, R. Zdunek, S. Amari. "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," *Independent Component Analysis and Blind Signal Separation*, pp. 32–39, 2006.

[9] C. Ding, T. Li, W. Peng. "Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method," *Proceedings of AAAI '06*, vol. 21, pp. 342–347, 2006.

[10] E. Gaussier, C. Goutte. "Relation between PLSA and NMF and implications," *Proceedings of ACM SIGIR*, pp. 601–602, 2005.

[11] T. Hofmann. "Probabilistic latent semantic indexing," *Proceedings of ACM SIGIR*, pp. 50–57, 1999.

[12] D.D. Lee, H.S. Seung. "Algorithms for non-negative matrix factorization," *Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[13] P. Sajda, S. Du, S., L. Parra. "Recovery of constituent spectra using non-negative matrix factorization," *Proc. of SPIE*, vol. 5207, pp. 321–331, 2003.

[14] J. Schler, M. Koppel, S. Argamon and J. Pennebaker. "Effects of Age and Gender on Blogging," *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006.

[15] Tibshirani, R. "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B*, pp. 267–288, 1996.

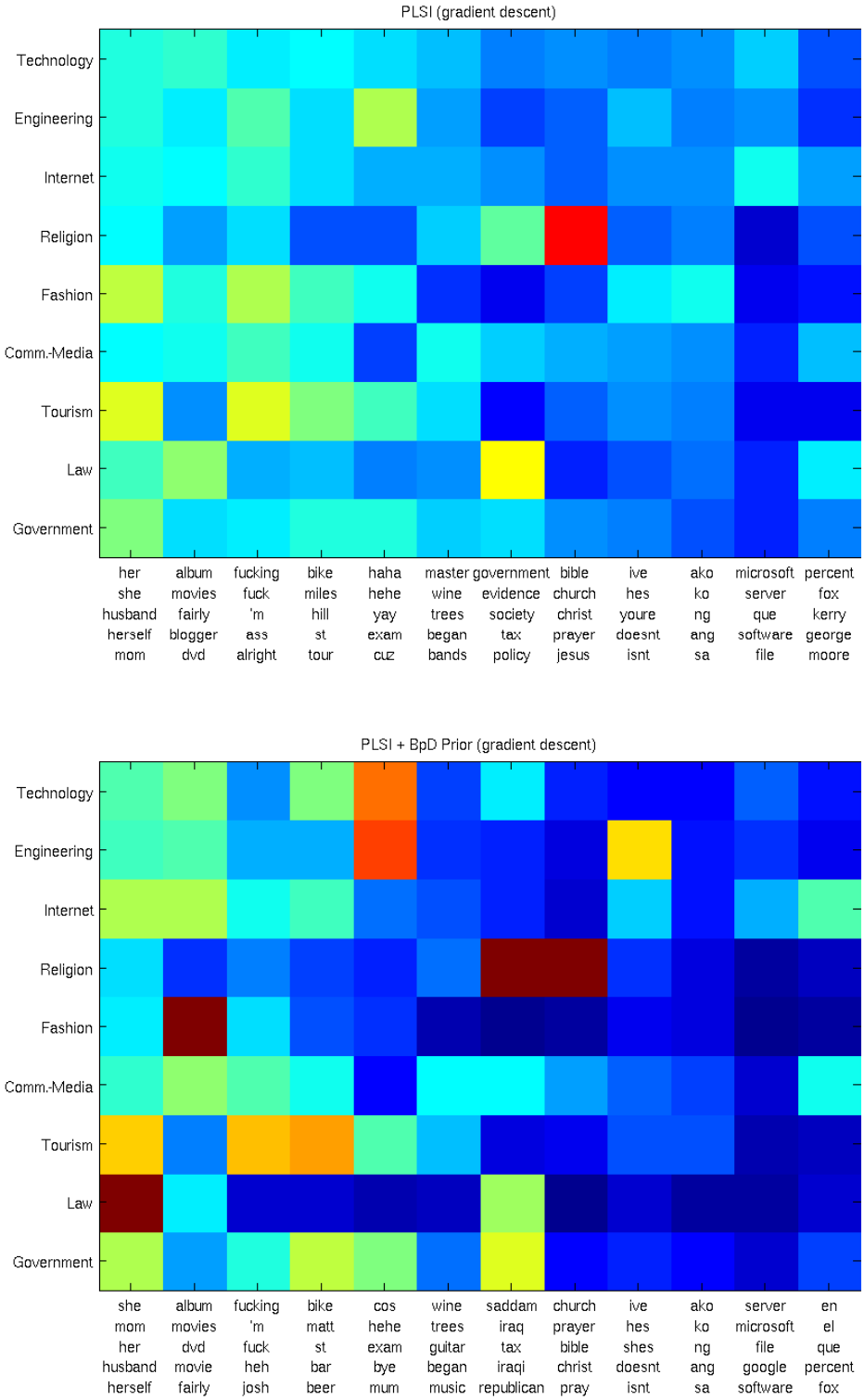

Figure 2: Comparison of ordinary PLSI to PLSI regularized with a Bounded pseudo-Dirichlet prior to induce sparsity, with the regularization parameter $\epsilon = 0.02$.