# Text Modeling with the Trace Norm

Jason D. M. Rennie
jrennie@gmail.com

April 14, 2006[*]

## 1 Introduction

We have two goals: (1) to find a low-dimensional representation of text that allows generalization to unseen data, and (2) to group documents according to their similarities. Clearly these tasks are related—similar documents can be represented more compactly than dissimilar documents and vice versa. First, we discuss a model for text that smoothly penalizes the rank of the parameter matrix. Then, we discuss how to apply this model to clustering.

## 2 Text Model

We use a generative model and assume that the documents are not identifiable beyond their term frequency statistics. Let $Y$ be the matrix of term frequency statistics, one document per row. Let $P(Y|X)$ be the likelihood model where $X$ is a matrix of parameters for the model. A simple example that we will use is a set of multinomial models, one per document where each row of $X$ parameterizes the multinomial for the corresponding document. We take the product of these per-document models to get $P(Y|X)$. We assume a prior on the parameter matrix, $P(X)$. This prior is one way in which we determine the similarity of documents, which, in turn, determines how compactly we can represent the documents. We make use of the trace norm (sum of singular values of a matrix) and discuss parameter and likelihood choices which are appropriate for its use.

### 2.1 Parameter Prior

The parameter prior for our model, $P(X)$, is the vehicle through which we obtain a reduced representation of the data. In particular, we want the prior to give high weight to a matrix that reflects natural similarities in the data. One natural bias is to weight highly parameter matrices of low rank. It is often the case that data is produced from relatively few underlying factors. This

[*] Updated June 16, 2006.

| $X_1$ | $X_2$ | $\|X\|_\Sigma$ | $\sin\theta$ |
|---|---|---|---|
| $(1,0)$ | $(1,0)$ | 1.41 | 0 |
| $(1,0)$ | $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ | 1.85 | .71 |
| $(1,0)$ | $(0,1)$ | 2 | 1 |
| $(1,0)$ | $\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ | 1.85 | .71 |
| $(1,0)$ | $(-1,0)$ | 1.41 | 0 |

Table 1: The first two columns give examples of pairs of unit length vectors. The third column gives the trace norm of the matrix of stacked vectors. The fourth column gives the sine of the angle between the vectors. The trace norm is closely approximated by a linear function of $\sin\theta$.

reflects the idea that documents might be composed of a combination of a few underlying themes or topics. This bias is common [2, 4, 1]. Another natural bias for text is to assume that the vectors corresponding to the data are close together. Term frequency vectors are often normalized to unit Euclidean length and treated as points on the unit sphere. In this representation, we can measure angles between vectors. Documents are similar if the angle between their vector representations are close. The trace norm can be used as a (smooth) measure of compactness of a set of vectors. It also has strong ties to matrix rank.

### 2.1.1 Trace Norm

The trace norm of a matrix is the sum of its singular values. Let $X = U\Sigma V^T$ be the singular value decomposition of $X$, where $U^T U = I$, $\Sigma = \text{diag}(\sigma)$, and $V^T V = I$. Then, the trace norm of $X$ is

$$\|X\|_\Sigma = \sum_i \sigma_i. \tag{1}$$

The trace norm is related to the rank of a matrix. In particular, Fazel (§5.1.4, 5.1.5 of [3]) showed that the convex hull of matrices with bounded rank $r$ is identical to the space of matrices with unit spectral norm and bounded trace norm ($\leq r$). Furthermore, the trace norm provides information about how "close" vectors are to each other.

To help the reader gain intuition for the trace norm, we provide simple examples in Table 1. Listed are pairs of 2-dimensional vectors, their trace norm when stacked as rows of a matrix, $X = [X_1; X_2]$, and the sine of the angle formed by the vectors (which, due to the simplicity of the example is just the second component of $X_2$). Unfortunately, the trace norm provides neither an upper nor lower bound on rank. Nor is it a linear function of the sine of the angle. However, it provides a smooth function of the matrix which is strongly correlated with these measures. In the example, the trace norm is largest when the vectors are perpendicular, and smallest when the vectors are co-linear. In our example, the trace norm is closely approximated by a linear function of the sine of the angle between vectors, $f(\theta) = a \sin\theta + b$ where $a = 2 - \sqrt{2}$ and $b = \sqrt{2}$.

Previous work on text modeling has focused on limiting the rank of the representation of the term frequency matrix. Here, we use the trace norm to give us a smooth penalty that encourages low rank and a compactness of the parameter vectors.

### 2.1.2 Trace Norm Prior

We utilize the trace norm by incorporating it into the parameter prior. We esablish a prior that is the exponentiated negative of the trace norm,

$$P(X) \propto \exp(-\|X\|_\Sigma). \tag{2}$$

Thus, maximization of a joint likelihood encourages a low-rank parameter matrix with row/column vectors in relatively close proximity. The negative log-likelihood (NLL) of the prior is the trace norm, so if we were to interpret our model as an encoding framework, the value of the trace norm would serve as the parameter encoding length [5].

## 2.2 Likelihood

The multinomial is a simple, standard model of term frequency for text. It is the model of term frequency that results from the unigram model, where the word for each position in the documents is drawn iid. The multinomial can be parameterized in multiple ways. Most common is the mean parameterization,

$$P(y|\theta) = \frac{n!}{\prod_i y_i!} \prod_i \theta_i^{y_i}, \tag{3}$$

where $n = \sum_i y_i$ is the fixed document length and the parameter for word $i$, $0 \le \theta_i \le 1$, is also its expected rate of occurrence. The above is for a single document, $y$, and parameter vector, $\theta$. For multiple documents, we take the product of likelihoods and stack parameters into a matrix, $P(Y|\Theta) = \prod_i P(Y_i|\Theta_i)$, where indices indicate row vectors. The mean parameterization is intuitively pleasing because each of the parameters $\{\theta_i\}$ is the expected rate of occurrence for the corresponding word. However, the fact that the parameters are constrained make it somewhat awkward to use with the trace norm prior. And, the logarithmic transform applied to the parameters in the NLL means that intuition is often wrong about the effects of small parameter changes. An alternate choice is the so-called "natural parameterization." The natural parameters are related to the mean parameters through a simple formula,

$$\theta_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}. \tag{4}$$

Unlike the mean paramterization, the natural parameters are unconstrained and their interaction with the data in the NLL is linear, so the effects of parameter changes are easier to understand. In the next two sections, we discuss the interactions of these parameterizations with the trace norm prior.

### 2.2.1 Mean Parameterization

For the mean parameterization, the likelihood values must be non-negative and sum-to-one. We can take care of this restriction via transformation between the parameters used for the prior, $X$, and those used for the likelihood, $\Theta$,

$$\Theta_{ij} = \frac{|X_{ij}|}{\sum_j |X_{ij}|}. \tag{5}$$

Thus, our joint probability becomes $P(Y|\Theta)P(X)$. But, in this context, the prior is meaningless since even as $\|X\|_\Sigma \to 0$ (equivalently, $P(X) \to 1$), we can attain the maximum likelihood. Even a simple constraint on the trace norm, such as $\|X\|_\Sigma \geq 1$, is insufficient to overcome the extra degrees of freedom introduced by the transform, (5). Why, then, don't we apply the trace norm prior directly to the mean parameters, $P(\Theta)$? The reason is that this would create an undesirable bias. The trace norm measures lengths using the $L_2$ norm. I.e. the trace norm of a mean parameter vector (which has unit $L_1$ length) can vary from 1 to $\frac{1}{\sqrt{d}}$, where $d$ is the dimensionality of the vector. Such a prior would serve to encourage low entropy parameter vectors more than it would encourage the set of parameter vectors to be of low rank. We can correct this bias by replacing the $L_1$ constraint with an $L_2$ constraint on the parameter vectors used for the trace norm calculation. Our joint probability is $P(Y|\Theta)P(X)$ where (5) is used to convert between $\Theta$ and $X$. For the optimization, we apply the constraint that each row of $X$ has unit $L_2$ norm. Thus, the trace norm of any row of $X$ is 1, and the prior only imparts a bias on the relative locations of the parameter vectors, not their absolute locations. The values in Table 1 give intuition for how the trace norm is affected by different orientations of a pair of unit $L_2$ length vectors.

### 2.2.2 Natural Parameterization

The natural parameterization leads to a simpler model. There is one undesirable bias. However, this is fixed naturally by introducing a parameter hierarchy. We begin with an short introduction to the exponential family of distributions.

A likelihood $P(y|x)$ is in the exponential family if it can be written as

$$P(y|x) = a(y)b(x)e^{c(y)^T d(x)}, \tag{6}$$

where $a$ and $b$ are functions that return scalar values; $c$ and $d$ are functions that return vector values. The density is in "canonical" form if $c(y) = y$; in canonical form, $d(x)$ is the natural paramter. For the multinomial, we have $a(y) = -\sum_i \log y_i!$, $b(x) = \log n! - n \log \sum_i \exp(x_i)$, $c(y) = y$ and $d(x) = x$, or

$$-\log P(y|x) = \sum_i \log y_i! - \log n! + n \log \sum_j \exp(x_j) - \sum_i y_i x_i, \tag{7}$$

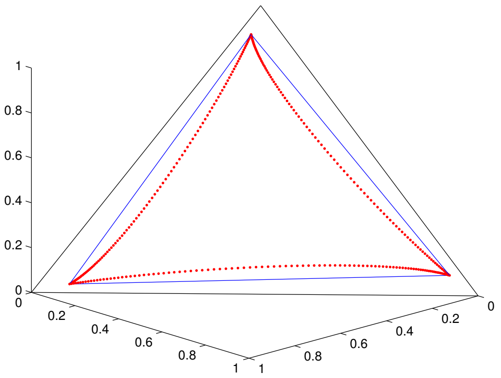

Figure 1: Shown are boundaries of regions enclosing the convex hull of the points (.9, .07, .03), (.07, .03, .9), and (.03, .9, .07). Black (outer, solid) lines bound the simplex. The blue (inner, solid) line bounds the mean parameter convex hull. The red (dotted) line bounds the natural parameter convex hull. To construct the natural parameter convex hull, we find natural parameter representations of the three points, determine the convex hull in natural parameter space, then transform the hull to mean parameter space. Viewed in mean parameter space, the natural parameter convex hull is not convex.

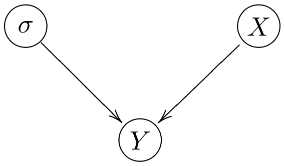

Figure 2: Graphical model for the multi-prior model.

where $n = \sum_i y_i$ is the document length. I.e. $P(y|x)$ is in canonical form and $x$ is the natural parameter. As with any natural parameterization, the rate of change in NLL for a small change in the parameter is the difference between expected and empirical values.

It is valuable to understand how natural parameters behave on the simplex. Due to the transformation between natural and mean parameters, (5), convex combinations of natural parmeters do not lead to convex regions on the simplex. Figure 1 provides an example.

Our natural parameter model is described by the joint distribution, $P(X)P(Y|X)$, where the likelihood is a product of natural parameter multinomials, $P(Y|X) = \prod_i P(Y_i|X_i)$, where $Y_i$ is the $i^{\text{th}}$ row of $Y$ and $P(X)$ is the trace norm prior. This model prefers a low magnitude, low rank parameter matrix. The low rank bias is desirable, but the low magnitude bias is not—it too strongly encourages points close to the middle of the simplex[1]. The trace norm prior penalizes each row of $X$ separately for its distance from the origin of the natural parameter space. The origin is the de facto "center" for the trace norm calculation. However, the origin is usually a poor center for text—some words are simply more common than others. We can shift the "center" by introducing a new parameter vector which is added to each row of $X$ for the likelihood parameters. Let $\sigma$ be this "center" parameter vector. Our updated joint distribution is $P(\sigma)P(X)\prod_i P(Y_i|X_i+\sigma)$. $X$ and $\sigma$ are independent for generation ($Y$ unobserved), but dependent for inference ($Y$ observed). We call this the "multi-prior" model (see Figure 2). A natural choice for $P(\sigma)$ is the trace norm prior[2]. In effect, only a single parameter vector will be biased towards the origin. Hence, for a large data set (many documents), this bias will be quite small.

Though unclear in the graphical model, the multi-prior model reflects hierarchical structure in the data. It can be easily extended to deal with additional structure, such as classes, sub-classes, etc.

# 3 Discussion

Use of the trace norm for a parameter prior for text modeling provides a smooth low-rank. The trace norm prior can be utilized in either a mean parameter or natural parameter framework. Each has minor issues that can be addressed easily, yielding two frameworks for learning low-rank text models. We find the natural parameter model particularly pleasing due to its intrinsic advantages—lack of constraints, a simple derivative, and ease in modeling of hierarchical structure.

# References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems 14*, 2002.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[3] M. Fazel. *Matrix Rank Minimization with Applications.* PhD thesis, Stanford University, 2002.

[4] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, 1999.

[5] J. D. M. Rennie. Encoding model parameters. people.csail.mit.edu/jrennie/writing, May 2003.

[1] The middle of the simplex corresponds to a natural parameter vector of all zeros.

[2] Note that since $\sigma$ is a vector, its trace norm is simply its $L_2$ length, $\|\sigma\|_\Sigma = \|\sigma\|_2$.