Sparse Additive Generative Models of Text

Jacob Eisenstein   JACOBEIS@CS.CMU.EDU
Amr Ahmed   AMAHMED@CS.CMU.EDU
Eric P. Xing   EPXING@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15203 USA

Abstract

Generative models of text typically associate a multinomial with every class label or topic. Even in simple models this requires the estimation of thousands of parameters; in multi-faceted latent variable models, standard approaches require additional latent "switching" variables for every token, complicating inference. In this paper, we propose an alternative generative model for text. The central idea is that each class label or latent topic is endowed with a model of the *deviation* in log-frequency from a constant background distribution. This approach has two key advantages: we can enforce sparsity to prevent overfitting, and we can combine generative facets through simple addition in log space, avoiding the need for latent switching variables. We demonstrate the applicability of this idea to a range of scenarios: classification, topic modeling, and more complex multifaceted generative models.

1. Introduction

Generative models of text overwhelmingly rely on the Dirichlet-multinomial conjugate pair. The primary advantage is that estimation is straightforward and efficient, with the Dirichlet prior contributing pseudo-counts to the observed counts generated by the multinomial. However, the ease of parameter estimation comes at a cost: unnecessarily complicated latent variable structures and lack of robustness to limited training data. More concretely, we see three main problems with Dirichlet-multinomial generative models:

- **Inference cost** There is increasing interest in modeling text with multiple generative facets, such as syntax (Griffiths et al., 2005), sentiment (Mei et al., 2007), and ideological and cultural perspective (Ahmed & Xing, 2010; Paul & Girju, 2010). In most cases, the incorporation of multiple facets requires an additional latent variable *per token*, to act as a "switch" controlling which facet is currently active. This huge number of additional latent variables makes inference more expensive.

- **Overparametrization** Standard Dirichlet-multinomial generative models learn a unique probability distribution over the entire vocabulary. General lexical patterns — for example, the high frequency of function words "the" and "of" — must be re-learned for every topic, wasting training data. In practice function words are typically removed using heuristics, or must be handled explicitly through additional latent variables (Chemudugunta et al., 2006).

- **Lack of sparsity** The Dirichlet-multinomial is incapable of using sparsity to limit model complexity. While the Dirichlet prior can induce zeros in the multinomials it generates, such sparsity is counterproductive to robustness: for example, supervised models that assign zero generative likelihood for some terms will be extremely brittle, because the label assignment of an entire document can be vetoed by a single word.

This paper proposes an alternative to the Dirichlet-multinomial for generative models of text: the Sparse Additive Generative model (SAGE). The problems with the Dirichlet-multinomial stem from a root cause: directly modeling the lexical probabilities associated with each document class or latent factor. In contrast, SAGE models the *difference* in log-frequencies from a background lexical distribution (see Figure 1). This has two key advantages: first, we can apply a
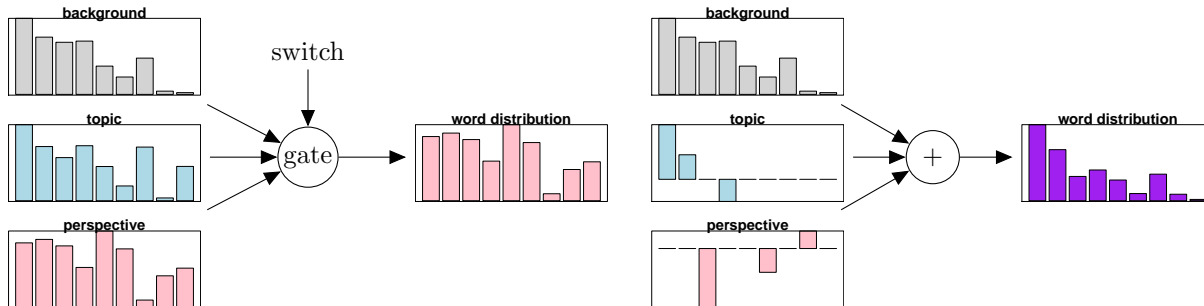
Figure 1. A schematic comparison between a standard multinomial switching model (left) and SAGE (right). Rather than choosing among probability distributions for each facet, SAGE additively combines sparse zero-mean variations.

sparsity-inducing prior to limit the number of terms whose probability diverges from the background lexical frequencies. This increases predictive accuracy and robustness to limited training data. Second, we can construct multi-faceted latent variable models by simply adding together SAGE component vectors. For example, if a blog post on the topic of climate change is written from a right-wing perspective, we can model the text by summing the SAGE components associated with the topic, perspective, and topic-perspective interaction. No latent "switching" variables are required to decide which of these components is active for a given token.

SAGE is intended as a drop-in replacement for the Dirichlet-multinomial, and can be applied in a broad range of generative models. We demonstrate SAGE's advantages in a number of different settings. First, we substitute SAGE for the Dirichlet-multinomial in a naïve Bayes text classifier, obtaining higher overall accuracy, especially in the face of limited training data. Second, we use SAGE in a topic model, obtaining better predictive likelihood on held-out text by learning simpler topics with less variation on rare words. Third, we apply SAGE in generative models which combine topics with additional facets: ideology and geographical variation.

2. Additive generative models of text

The core idea of our generative model is that of a background lexical distribution, which is modified by adding additional vectors in log-space. In the simplest case, we have a background distribution $\boldsymbol{m} \in \mathcal{R}^V$, and a set of components $\{\boldsymbol{\eta}_k \in \mathcal{R}^V\}$, where V is the size of the vocabulary. Each component η_k corresponds to a document label $y \in 1 \ldots Y_{\max}$. Then the generative

distribution for each word in a document d is,

$$P(w|y_d, \boldsymbol{m}, \boldsymbol{\eta}) = \frac{\exp\left(\boldsymbol{m} + \boldsymbol{\eta}_{y_d}\right)}{\sum_i \exp\left(m_i + \eta_{y_d,i}\right)} \tag{1}$$

In this formulation each document has a single component index y_d. If this index is observed then this model corresponds to a naïve Bayes model of text, where we have substituted the vector of log frequency deviations $\boldsymbol{\eta}$ for the standard multinomial; if y_d is not observed then the model is a mixture of unigrams. We will extend this framework to include per-word latent indices, drawn from a document-specific topic distribution — this corresponds to a variant of latent Dirichlet allocation (Blei et al., 2003), in which each topic is represented by log frequency deviations $\boldsymbol{\eta}$ rather than word probabilities $\boldsymbol{\beta}$. Thus, we can replicate the most common existing generative models of text using SAGE, taking advantage of sparsity-inducing priors on $\boldsymbol{\eta}$ to obtain additional robustness.

However, SAGE has another advantage: by modeling log term frequencies rather than raw probabilities, it is easy to combine multiple facets simply by adding them. We can replicate existing models by adding a sparse deviation vector $\boldsymbol{\eta}$ to the background log-frequencies $\boldsymbol{m}$; by adding additional facets, we obtain richer and more complex models. For example, in a topic-perspective model, we associate a single perspective y_d with the entire document; for each token w_n we have a unique topic z_n. Using Dirichlet-multinomials would require an additional switching variable to determine whether the token w_n is drawn from the topic $\boldsymbol{\beta}_{z_n}^{(t)}$ or the perspective $\boldsymbol{\beta}_{y_d}^{(p)}$ (Ahmed & Xing, 2010). But with SAGE, we draw the token w_n from a distribution proportional to $\exp\left(\boldsymbol{m} + \boldsymbol{\eta}_{y_d}^{(p)} + \boldsymbol{\eta}_{z_n}^{(t)}\right)$, without need for latent switching variables.

We ignore covariance between terms and treat each element η_{ki} independently, where k indexes the com-

ponent vector $\boldsymbol{\eta}_k$ and i indexes into the vocabulary. A zero-mean Laplace prior has the same effect as placing an L_1 regularizer on η_{ki}, inducing sparsity while at the same time permitting more extreme deviations from the mean. The Laplace distribution $\mathcal{L}(\eta; m, \sigma)$ is equivalent to a compound model, $\int \mathcal{N}(\eta; m, \tau)\mathcal{E}(\tau; \sigma)d\tau$, where $\mathcal{E}(\tau; \sigma)$ indicates the Exponential distribution (Lange & Sinsheimer, 1993; Figueiredo, 2003). This identity is the cornerstone of our inference, which treats the variance τ as a latent variable. We now present a generative story for the incorporation of SAGE in a naïve Bayes classifier:

- Draw the background $\boldsymbol{m}$ from an uninformative prior
- For each class k
  - For each term i
    * Draw $\tau_{k,i} \sim \mathcal{E}(\gamma)$
    * Draw $\eta_{k,i} \sim \mathcal{N}(0, \tau_{k,i})$
  - Set $\boldsymbol{\beta}_k \propto \exp(\boldsymbol{\eta}_k + \boldsymbol{m})$
- For each document d
  - Draw a class y_d from a uniform distribution
  - For each word n, draw $w_n^{(d)} \sim \boldsymbol{\beta}_{y_d}$

In general we work in a Bayesian setting, but for the components η we take maximum *a posteriori* point estimates. Bayesian uncertainty is problematic due to the logistic transformation: even if the expectation $\langle \eta_{ki} \rangle = 0$, any posterior variance over η_{ki} would make $\langle \exp(\eta_{ki} + m_i) \rangle > \langle \exp m_i \rangle$. We resort to a combination of MAP estimation over η and Bayesian inference over all other latent variables — this is similar to the treatment of the topics $\boldsymbol{\beta}$ in the original formulation of latent Dirichlet allocation (Blei et al., 2003). The background word distribution $\boldsymbol{m}$ is assumed to be fixed, and we fit a variational distribution over the remaining latent variables, optimizing the bound,

$$\ell = \sum_d \sum_n^{N_d} \log P(w_n^{(d)} | \boldsymbol{m}, \boldsymbol{\eta}_{y_d}) + \sum_k \langle \log P(\boldsymbol{\eta}_k | \mathbf{0}, \boldsymbol{\tau}_k) \rangle$$
$$+ \sum_k \langle \log P(\boldsymbol{\tau}_k | \gamma) \rangle - \sum_k \langle \log Q(\boldsymbol{\tau}_k) \rangle, \quad (2)$$

where N_d is the number of tokens in document d.

3. Estimation

We now describe how SAGE components can be efficiently estimated using a Newton optimization.

3.1. Component means

First we address learning the component vectors $\boldsymbol{\eta}$. Letting $\boldsymbol{c}_d$ represent the vector of term counts for document d, and $C_d = \sum_i \boldsymbol{c}_{di}$, we compute the relevant

parts of the bound,

$$\ell(\boldsymbol{\eta}_k) = \sum_{d:c_d=k} \boldsymbol{c}_d^\mathsf{T} \boldsymbol{\eta}_k - C_d \log \sum_i \exp(\eta_{ki} + m_i)$$
$$- \boldsymbol{\eta}_k^\mathsf{T} \text{diag}\left(\langle \boldsymbol{\tau}_k^{-1} \rangle\right) \boldsymbol{\eta}_k / 2 \quad (3)$$

$$\frac{\partial \ell}{\partial \boldsymbol{\eta}_k} = \boldsymbol{c}_k - C_k \frac{\exp(\boldsymbol{\eta}_k + \boldsymbol{m})}{\sum_i \exp(\eta_{ki} + m_i)} - \text{diag}\left(\langle \boldsymbol{\tau}_k^{-1} \rangle\right) \boldsymbol{\eta}_k$$
$$= \boldsymbol{c}_k - C_k \boldsymbol{\beta}_k - \text{diag}\left(\langle \boldsymbol{\tau}_k^{-1} \rangle\right) \boldsymbol{\eta}_k, \quad (4)$$

abusing notation so that $\boldsymbol{c}_k = \sum_{d:c_d=k} \boldsymbol{c}_d$ and $C_k = \sum_i \boldsymbol{c}_{ki}$. Note that the fraction $\frac{\exp(\boldsymbol{\eta}_k + \boldsymbol{m})}{\sum_i \exp(\eta_{ki} + m_i)}$ is equal to the term frequency vector $\boldsymbol{\beta}_k$. The gradient has an intuitive interpretation as the difference of the true counts $\boldsymbol{c}_k$ from their expectation $C_k \boldsymbol{\beta}_k$, minus the divergence of $\boldsymbol{\eta}$ from its prior mean $\mathbf{0}$, scaled by the expected inverse-variance. We will use Newton's method to optimize $\boldsymbol{\eta}$, so we first derive the Hessian,

$$\frac{d^2\ell}{d\eta_{ki}^2} = C_k \beta_{ki}(\beta_{ki} - 1) - \langle \tau_{ki}^{-1} \rangle, \frac{d^2\ell}{d\eta_{ki}d\eta_{ki'}} = C_k \beta_{ki}\beta_{ki'}$$
$$\mathbf{H}(\boldsymbol{\eta}_k) = C_k \boldsymbol{\beta}_k \boldsymbol{\beta}_k^\mathsf{T} - \text{diag}\left(C_k \beta_k + \langle \boldsymbol{\tau}_k^{-1} \rangle\right). \quad (5)$$

The Hessian matrix $\mathbf{H}$ is rank-one plus diagonal, so it can be efficiently inverted using the Sherman-Morrison formula. For notational simplicity, we elide the class index k, and define the convenience variable $\mathbf{A} = \text{diag}\left(-(C\boldsymbol{\beta} + \langle \boldsymbol{\tau}^{-1} \rangle)^{-1}\right)$. We can now derive a Newton optimization step for $\boldsymbol{\eta}$, using the gradient $\boldsymbol{g}(\boldsymbol{\eta}) = \frac{\partial \ell}{\partial \boldsymbol{\eta}}$ from Equation 4:

$$\mathbf{H}^{-1}(\boldsymbol{\eta}) = \mathbf{A} - \frac{\mathbf{A}C\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}\mathbf{A}}{1 + C\boldsymbol{\beta}^\mathsf{T}\mathbf{A}\boldsymbol{\beta}}$$
$$-\Delta\boldsymbol{\eta} = \mathbf{H}^{-1}(\boldsymbol{\eta})\boldsymbol{g}(\boldsymbol{\eta})$$
$$= \mathbf{A}\boldsymbol{g}(\boldsymbol{\eta}) - \frac{C\mathbf{A}\boldsymbol{\beta}}{1 + C\boldsymbol{\beta}^\mathsf{T}\mathbf{A}\boldsymbol{\beta}}\left[\boldsymbol{\beta}^\mathsf{T}(\mathbf{A}\boldsymbol{g}(\boldsymbol{\eta}))\right], \quad (6)$$

where the parenthesization defines an order of operations that avoids forming any non-diagonal matrices. Thus, the complexity of each Newton step is linear in the size of the vocabulary.

3.2. Variances

Next we consider the variance; recall that we have a random vector $\boldsymbol{\tau}_k$ for every component k. Unlike the components $\boldsymbol{\eta}$, we are Bayesian with respect to $\boldsymbol{\tau}$, and construct a fully-factored variational distribution $Q_{\boldsymbol{\tau}_k}(\boldsymbol{\tau}_k) = \prod_i Q_{\tau_{ki}}(\tau_{ki})$. We set the form $Q_{\tau_{ki}}$ to be a Gamma distribution with parameters $\langle a, b \rangle$:

$$Q(\tau) = \mathcal{G}(\tau; a, b) = \tau^{a-1}\frac{\exp(-\tau/b)}{\Gamma(a)b^a},$$

so that $\langle \tau \rangle = ab, \langle \tau^{-1} \rangle = ((a-1)b)^{-1}$, and $\langle \log \tau \rangle = \psi(a) + \log(b)$. The prior on τ is an Exponential

distribution with parameter γ, so that $P(\tau|\gamma) = \gamma \exp(-\gamma\tau)$. We can now define the contribution to the variational bound from $Q(\tau)$,

$$\ell(\tau) = \langle \log P(\eta|\tau) \rangle + \langle \log P(\tau|\gamma) \rangle - \langle \log Q(\tau) \rangle$$

$$\propto -\frac{1}{2} \langle \log \tau \rangle - \frac{1}{2}\eta^2 \langle \tau^{-1} \rangle - \gamma \langle \tau \rangle$$

$$- (a-1) \langle \log \tau \rangle + \langle \tau \rangle /b + \log \Gamma(a) + a \log b \quad (7)$$

$$-\Delta a = \frac{(1/2-a)\psi_1(a) + \frac{1}{2}\eta^2 b^{-1}(a-1)^{-2} - \gamma b + 1}{(1/2-a)\psi_2(a) - \psi_1(a) - \eta^2 b^{-1}(a-1)^{-3}}$$

$$b = \frac{1 + \sqrt{1 + 8\gamma\eta^2 \frac{a}{a-1}}}{4\gamma a},$$

obtaining a Newton optimization for a and a closed-form update for b. We use $\psi_1(a)$ and $\psi_2(a)$ indicate the trigamma and quad-gamma functions respectively.

For a parameter-free model, we can replace the Exponential prior on τ with an improper Jeffrey's prior, $P(\tau) \propto 1/\tau$. The combination of the Jeffrey's prior $P(\tau)$ with the Gaussian $P(\eta|0,\tau)$ no longer yields a Laplace distribution. However, the Normal-Jeffrey's compound distribution also induces sparsity, and relieves us from having to choose a value for γ; moreover, Guan & Dy (2009) find that it yields better results for sparse probabilistic PCA than the Laplace distribution. To derive the variational parameters of $Q(\tau)$, we need only replace the term $-\gamma\langle\tau\rangle$ in Equation 7 with $-\langle \log \tau \rangle$. The resulting updates are,

$$-\Delta a = \frac{(1/2+a)\psi_1(a) - \frac{1}{2}\eta^2 b^{-1}(a-1)^{-2} - 1}{(1/2+a)\psi_2(a) + \eta^2 b^{-1}(a-1)^{-3}}$$

$$b = \frac{\eta^2}{a-1}.$$

These variational parameters are necessary only to compute the bound; we can directly compute the expectation $\langle \tau_{ki}^{-1} \rangle = \eta_{ki}^2$.

Application 1: Document classification

Our first evaluation is on document classification: we test SAGE as a drop-in replacement for the multinomial-Dirichlet that is traditionally used in naïve Bayes text classifiers. Both generative models are parameter-free: for SAGE, we use the non-parametric Jeffrey's prior on the variance τ; for the multinomial-Dirichlet, we perform a coordinate ascent in which a Newton optimization (Minka, 2003) is used to update the precision of the Dirichlet prior.[1] Discriminative methods may yield better performance on the document classification task, but our goal here is

---

[1]Wallach et al (2009a) find that asymmetric Dirichlet priors for term distributions provide no advantage over symmetric priors.
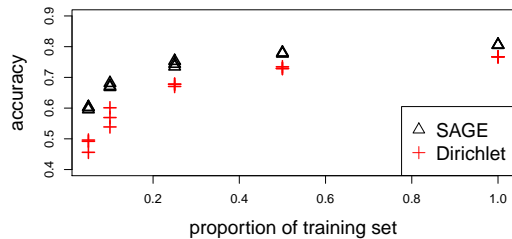

Figure 2. Accuracy on 20 Newsgroup text classification, as the amount of training data is varied

to compare generative models which are amenable to incorporation in the more complex latent variable settings described in the remainder of the paper.

We evaluate on the classic benchmark "Twenty Newsgroups" data, in which the task is to classify unlabeled newsgroup postings into twenty different newsgroups. Using the training/test split from the website http://people.csail.mit.edu/jrennie/20Newsgroups/, there are 11,269 training documents; we randomly subsampled a range of training sets including as few as 5% of the original documents. We did not perform stopword filtering, and used a vocabulary of 50,000 terms. Results are shown in Figure 2: the amount of training data varied on the x-axis; each point corresponds to a different random subsample. SAGE substantially outperforms the Dirichlet-multinomial in every experiment. Its advantage is particularly robust in the limited-data settings, where the raw improvement in accuracy is more than 10%. The Jeffrey's prior on τ adaptively controls the sparsity, which increases monotonically from 90% in the full-data setting to more than 98% in the minimal data setting. Performance gains were smaller in a pilot experiment with a vocabulary of 10,000, suggesting that SAGE's strength is its ability to exploit rare words without overfitting.

4. Latent variable models

Next, we consider how to incorporate SAGE in a latent variable model of text. We focus on topic models, which contain one latent discrete variable per token, and a latent vector of topic proportions per document. The generative story is similar to the document classification model the previous section, with the following additions: each document is endowed with a vector of topic proportions $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha)$; each token has an associated latent topic $z_n^{(d)}$; and the probability distri-

bution for a given token is

$$P(w_n^{(d)}|z_n^{(d)}, \boldsymbol{\eta}, \boldsymbol{m}) \propto \exp\left(\boldsymbol{\eta}_{z_n^{(d)}} + \boldsymbol{m}\right).$$

We can combine the mean field variational inference for latent Dirichlet allocation (LDA) with the variational treatment of τ, optimizing the bound,

$$\ell = \sum_d \langle \log P(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \rangle + \sum_n^{N_d} \left\langle \log P(w_n^{(d)}|\boldsymbol{m}, \boldsymbol{\eta}_{z_n^{(d)}}) \right\rangle$$
$$+ \left\langle \log P(z_n^{(d)}|\boldsymbol{\theta}_d) \right\rangle + \sum_k \langle \log P(\boldsymbol{\eta}_k|\boldsymbol{0}, \boldsymbol{\tau}_k) \rangle$$
$$+ \sum_k \langle \log P(\boldsymbol{\tau}_k|\gamma) \rangle - \langle \log Q(\boldsymbol{\tau}, \boldsymbol{z}, \boldsymbol{\theta}) \rangle . \quad (8)$$

The updates for $Q(z)$ and $Q(\boldsymbol{\theta})$ are identical to standard LDA; the updates for $Q(\tau)$ remain as Section 3.2. However, the presence of latent variables slightly changes the MAP estimation for $\boldsymbol{\eta}$:

$$\ell(\boldsymbol{\eta}_k) = \sum_d \sum_n^{N_d} Q_{z_n^{(d)}}(k) \left(\boldsymbol{\eta}_k - \log \sum_i \exp\left(\eta_{ki} + m_i\right)\right)$$
$$- \boldsymbol{\eta}_k^{\mathsf{T}} \mathrm{diag}\left(\langle \boldsymbol{\tau}_k^{-1} \rangle\right) \boldsymbol{\eta}_k/2$$
$$\frac{\partial \ell}{\partial \boldsymbol{\eta}_k} = \langle \boldsymbol{c}_k \rangle - \langle C_k \rangle \frac{\exp\left(\boldsymbol{\eta}_k + \boldsymbol{m}\right)}{\sum_i \exp\left(\eta_{ki} + m_i\right)} - \mathrm{diag}\left(\langle \boldsymbol{\tau}_k^{-1} \rangle\right) \boldsymbol{\eta}_k,$$

where $\langle c_{ki} \rangle = \sum_d \sum_n Q_{z_n^{(d)}}(k)\delta(w_n^{(d)} = i)$, and $\langle C_k \rangle = \sum_i \langle c_{ki} \rangle$. Thus, the exact counts $\boldsymbol{c}_k$ are replaced with their expectations under $Q(z)$.

We define an EM procedure in which the M-step consists in iteratively fitting the parameters $\boldsymbol{\eta}$ and $Q(\boldsymbol{\tau})$. It is tempting to perform a "warm start" by intializing with the values from a previous iteration of the outer EM loop. However, these parameters are tightly coupled: as the component mean η_{ki} goes to zero, the expected variance $\langle \tau_{ki} \rangle$ is also driven to zero; once $\langle \tau_{ki} \rangle$ is very small, η_{ki} cannot move away from zero regardless of the expected counts $\boldsymbol{c}_k$. This means that a warm start risks locking in a sparsity pattern during the early stages of EM which may be far from the global optimum. There are two solutions: either we abandon the warm start (thus expending more computation), or we do not iterate to convergence in each M-step (thus obtaining noisier and less sparse solutions, initially). Fortunately, pilot experiments showed that good results can be obtained by performing just one iteration in each M-step, while using the warm start technique.

Application 2: Sparse topic models

Our second evaluation applies the SAGE Topic Model to the benchmark NIPS dataset.[2] Following the evaluation of Wang and Blei (2009), we subsample to 10%

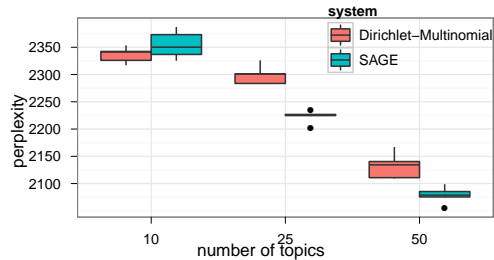[2]http://www.cs.nyu.edu/~roweis/data.html

Figure 3. Perplexity results for SAGE and latent Dirichlet allocation on the NIPS dataset.
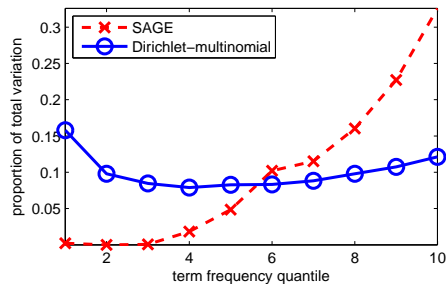

Figure 4. Proportion of total variation committed to words at each frequency decile. Dirichlet-multinomial LDA makes large distinctions in the topic-term frequencies of very rare words, while SAGE only distinguishes the topic-term frequencies of words with robust counts.

of the tokens in each document, and hold out 20% of the documents as a "test set" on which to evaluate predictive likelihood. We limit the vocabulary to the 10,000 terms that appear in the greatest number of documents; no stopword pruning is applied. Overall this leaves 1986 training documents with 237,691 tokens, and a test set of 497 documents and 57,427 tokens. We evaluate perplexity using the Chib-style estimation procedure of Wallach et al. (2009b). For comparison, we implement variational latent Dirichlet allocation, making maximum-likelihood updates to a symmetric Dirichlet prior on the topic-term distributions.

Results are shown in Figure 3, using box plots over five paired random initializations for each method. SAGE outperforms standard latent Dirichlet allocation as the number of topics increases; with both 25 and 50 topics, every SAGE run outperformed its counterpart Dirichlet-multinomial. As in the classification task, SAGE controls sparsity adaptively: as the number of topics increases from 10 to 50, the proportion of non-zero weights decreased five-fold (from 5% to 1%), holding the total model complexity constant.

Figure 4 compares the overall term frequency (measured in deciles and shown on the x-axis) with the amount of topic-term variation that each model accords, for a single run with 50 topics. In SAGE, the total variation for a term i is $\sum_k |\eta_{ki}|$, while in Dirichlet-multinomial LDA, we measure the total variation from the mean log frequency, $\sum_k |\beta_{ki} - \overline{\beta}_i|$. The figure shows that SAGE admits very little topical variation for low frequency words. In contrast, the Dirichlet-multinomial displays little sensitivity to the overall term frequency, and actually assigns *more* topical variation to the lowest frequency terms. Note that our implementation of Dirichlet-multinomial LDA incorporates a symmetric Dirichlet prior which acts to smooth the topic-term frequencies of rare words — even so, it overfits the training data and learns widely divergent probabilities for these words. We believe that this phenomenon explains the better predictive performance obtained by the SAGE topic model — by focusing on high-frequency terms with accurate counts, it learns more robust topics. A related point is that the topics induced by standard LDA may be more difficult to interpret, because the rare words may cause documents to be assigned to topics in a way that is not predictable from simply examining the most salient terms in each topic.

5. Multifaceted generative models

Finally, we consider how SAGE can be used to combine multiple generative facets. We focus on models that combine per-word latent topics and document labels, thus offering a structured view of labels and topics — for example, revealing the words and documents that reflect a left-wing perspective on education policy. Existing multifaceted generative models (Mei et al., 2007; Paul & Girju, 2010; Ahmed & Xing, 2010) incorporate latent "switch" variables that determine whether each word token is generated from a topic or from a distribution associated with the document label (as in the left panel of Figure 1). If the token is to be drawn from a topic, then an *additional* latent variable determines which topic will be active. Topic-label interactions can also be included, capturing the distributions of words at the intersection of, say, topic and ideology. The number of parameters thus becomes very large, growing to the product of the vocabulary size, the number of topics, and the number of labels — plus the additional switching variable per token. Collapsed Gibbs sampling can analytically marginalize the topic and label word distributions, but it still may suffer from high variance if the number of parameters is too large compared to the training data.
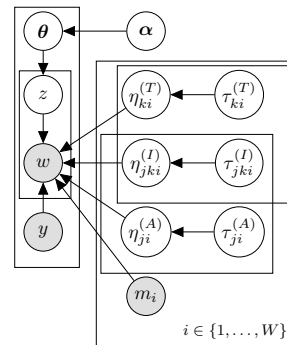

Figure 5. Plate diagram for a multifaceted topic model using SAGE.

SAGE enables multifaceted topic models that encourage sparse variation from the background term distribution, while eliminating the need for switching variables.[3] A plate diagram is shown in Figure 5. On the left, we have the standard document plate from latent Dirichlet allocation, augmented with the observed label y. On the right, we have an outer plate that makes explicit the repetition across all W words in the vocabulary. Within this plate, we have the observed background term frequency m_i, as well as sparse deviations for: each topic $\eta_{ki}^{(T)}$, $k < K$; each label distribution $\eta_{ji}^{(A)}$, $j < A$; and each topic-label interaction, $\eta_{jki}^{(I)}$. The variance parameters from the compound Normal-Jeffrey's distribution are shown as $\tau_{ki}^{(T)}$, etc. The generative probability for a single token is obtained by adding the SAGE components to the prior term frequencies:

$$P(w_n^{(d)}|z_n^{(d)},\boldsymbol{\eta},\boldsymbol{m},y_d) \propto \exp\left(\boldsymbol{\eta}_{z_n^{(d)}}^{(T)} + \boldsymbol{\eta}_{y_d}^{(A)} + \boldsymbol{\eta}_{y_d,z_n^{(d)}}^{(I)} + \boldsymbol{m}\right).$$

Estimation is very similar to the models encountered earlier in the paper. For the topic components $\boldsymbol{\eta}^{(T)}$, we obtain the gradient,

$$\frac{\partial\ell}{\partial\boldsymbol{\eta}_k^{(T)}} = \left\langle \boldsymbol{c}_k^{(T)}\right\rangle - \sum_j \langle C_{jk}\rangle \boldsymbol{\beta}_{jk} - \mathrm{diag}(\langle(\boldsymbol{\tau}_k^{(T)})^{-1}\rangle)\boldsymbol{\eta}_k^{(T)},$$

where $\boldsymbol{\beta}_{jk} \propto \exp\left(\boldsymbol{\eta}_k^{(T)} + \boldsymbol{\eta}_j^{(A)} + \boldsymbol{\eta}_{jk}^{(I)} + \boldsymbol{m}\right)$ and $\langle C_{jk}\rangle$ gives the expected counts for each topic-label combination. Using this gradient, we can apply the Newton optimization as before, substituting $\sum_j C_{jk}\boldsymbol{\beta}_{jk}$ into the Hessian in place of $C_k\boldsymbol{\beta}_k$ (see Equations 5 and 6).

The updates for $\boldsymbol{\eta}^{(A)}$ are almost identical, but we have exact counts $C_j^{(A)}$, as the labels are observed. The in-

---

[3] Zhu et al. (2006) also augment LDA topics using addition of log term frequencies (for per-token labels), but they did not employ sparsity.
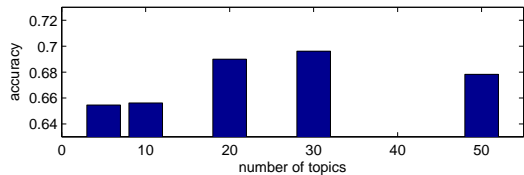
Figure 6. SAGE's accuracy on the ideological perspective task; the state-of-the-art is 69.1% (Ahmed & Xing, 2010).

teraction components $\boldsymbol{\eta}_{jk}^{(I)}$ depend on exactly one label distribution $\boldsymbol{\eta}_j^{(A)}$ and one topic $\boldsymbol{\eta}_k^{(T)}$, so we can use $C_{jk}\boldsymbol{\beta}_{jk}$ directly without computing any sums.

Application 3: Topic and ideology

We first evaluate on a publicly-available dataset of political blogs describing the 2008 U.S. presidential election (Eisenstein & Xing, 2010). There are a total of six blogs — three from the right and three from left — comprising 20,827 documents, 5.1 million word tokens, and a vocabulary of 8284 items. The task is to predict the ideological perspective of two unlabeled blogs, using the remaining four as a training set. We strictly follow the experimental procedure of Ahmed & Xing (2010), allowing us to compare with their reported results directly.[4]

Ahmed and Xing considered three ideology-prediction tasks, and found that the six-blog task was the most difficult: their Multiview LDA model achieves accuracy between 65% and 69.1% depending on the number of topics. They find a comparable result of 69% using support vector machines; alternative latent variable models Discriminative LDA (Lacoste-Julien et al., 2008) and Supervised LDA (Blei & McAuliffe, 2007) do worse. Our results are shown in Figure 6, reporting the median across five random initialization at each setting for K. Our best median result is 69.6%, at $K = 30$, equalling the state-of-the-art; our best individual run achieves 71.9%. Our model obtains sparsity of 93% for topics, 82% for labels, and 99.3% for topic-label interactions; nonetheless, pilot experiments show that the absence of topic-label interactions reduces performance substantially.

Application 4: Geolocation from text

We now consider the setting in which the label is itself a latent variable, generating both the text (as described above) and some additional metadata. This is the setting for the Geographical Topic Model, in which a latent "region" helps to select the distributions that

---

[4]The sole exception is that we learn the prior α from data, while Ahmed & Xing set it manually.

Table 1. Prediction error for Twitter geolocation.

error in kilometers:	median	mean
(Eisenstein et al., 2010)	494	900
(Wing & Baldridge, 2011)	**479**	967
SAGE	501	**845**

generate both text and observed GPS locations (Eisenstein et al., 2010). By training on labeled examples in which both text and geolocation are observed, the model is able to make predictions about the GPS location of unlabeled authors.

The Geographic Topic Model induces region-specific versions of each topic by chaining together log-Normal distributions. This is equivalent to an additive model in which both the topic and the topic-region interaction exert a zero-mean Gaussian deviation from a background language model. SAGE differs by allowing effects that are region-specific but topic-neutral, and by inducing sparsity. We follow the tuning procedures from Eisenstein et al. (2010) exactly: the number of latent regions is determined by running a Dirichlet process mixture model on the location data alone, and the number of topics is tuned against a development set. We also present more recent results from Wing & Baldridge (2011), who use a nearly identical dataset, but include a larger vocabulary. As shown in Table 1, SAGE achieves the best mean error of any system on this task, though Wing & Baldridge (2011) have the best median error.

6. Related work

Sparse learning (Tibshirani, 1996; Tipping, 2001) typically focuses on supervised settings, learning a sparse set of weights that minimize a loss on the training labels. Two recent papers apply sparsity to topic models. Williamson et al. (2010) induce sparsity in the topic proportions by using the Indian Buffet Process to represent the presence or absence of topics in a document. More closely related is the SparseTM of Wang & Blei (2009), which induces sparsity in the topics themselves using a spike-and-slab distribution. However, the notion of sparsity is different: in the SparseTM, the topic-term probability distributions can contain zeros, while in SAGE, each topic is a set of sparse deviations from a background distribution. Inference in the SparseTM requires computing a combinatorial sum over all sparsity patterns, while in our case a relatively simple coordinate ascent is possible.

Sparse dictionary learning provides an alternative approach to modeling document content with sparse bases (Jenatton et al., 2010). In general, such ap-

proaches emphasize sparsity in the number of dictionary components that are active for a given document. However, the application of a sparsity-inducing prior to the dictionary components would be similar to SAGE. The fundamental difference is that SAGE is a generative model that defines the probability of each token; as such, it can easily be embedded in larger latent variable structures.

7. Conclusion

We have presented SAGE, a new generative model for discrete data. Each token is generated by adding background log-probabilities to a set of sparse variation vectors associated with each generative factor. Applying SAGE to naïve Bayes classification and topic modeling, we find that it learns simpler models with better predictive performance. We feel that the most promising feature of SAGE is its facilitation of the construction of multifaceted generative models. We plan to explore the application of SAGE to even richer multifaceted generative models, such as hierarchical topic-aspect models and mixed-effects models that account for author-specific linguistic patterns.

References

Ahmed, Amr and Xing, Eric P. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of EMNLP*, pp. 1140–1150, 2010.

Blei, D. M. and McAuliffe, J. D. Supervised topic models. In *NIPS*, 2007.

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

Chemudugunta, Chaitanya, Smyth, Padhraic, and Steyvers, Mark. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In *Advances in Neural Information Processing Systems*, 2006.

Eisenstein, Jacob and Xing, Eric. The CMU 2008 political blog corpus. Technical report, Carnegie Mellon University, 2010.

Eisenstein, Jacob, O'Connor, Brendan, Smith, Noah A., and Xing, Eric P. A latent variable model of geographic lexical variation. In *Proceedings of EMNLP*, 2010.

Figueiredo, Mário A. T. Adaptive sparseness for supervised learning. *Pattern Analysis and Machine Learning*, 2003.

Griffiths, Thomas L., Steyvers, Mark, Blei, David M., and Tenenbaum, Joshua B. Integrating topics and syntax. In *Neural Information Processing Systems*, pp. 537–544, 2005.

Guan, Yue and Dy, Jennifer G. Sparse probabilistic principal component analysis. In *Proceedings of AISTATS*, 2009.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of ICML*, 2010.

Lacoste-Julien, S., Sha, F., and Jordan, M. I. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, 2008.

Lange, Kenneth and Sinsheimer, Janet S. Normal/Independent Distributions and Their Applications in Robust Regression. *Journal of Computational and Graphical Statistics*, 2(2), 1993.

Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. X. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of WWW*, 2007.

Minka, T. P. Estimating a Dirichlet distribution. Technical report, Massachusetts Institute of Technology, 2003.

Paul, M. and Girju, R. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of AAAI*, 2010.

Tibshirani, Robert. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Tipping, Michael E. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

Wallach, Hanna M., Mimno, David, and McCallum, Andrew. Rethinking LDA: Why Priors Matter. In *Neural Information Processing Systems*, 2009a.

Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan, and Mimno, David. Evaluation methods for topic models. In *Proceedings of ICML*, pp. 1105–1112, 2009b.

Wang, Chong and Blei, David M. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Neural Information Processing Systems*, 2009.

Williamson, S., Wang, C., Heller, K., and Blei, D. The IBP compound dirichlet process and its application to focused topic modeling. In *Proceedings of ICML*, 2010.

Wing, Benjamin and Baldridge, Jason. Simple supervised document geolocation with geodesic grids. In *Proceedings of ACL*, 2011.

Zhu, Xiaojin, Blei, David M., and Lafferty, John. Taglda: Bringing document structure knowledge into topic models. Technical Report TR-1553, University of Wisconsin, 2006.