# Class-Specific Simplex-Latent Dirichlet Allocation for Image Classification

Mandar Dixit[1,*] Nikhil Rasiwasia[2,*] Nuno Vasconcelos[1]
[1]Department of Electrical and Computer Engineering
University of California, San Diego
[2]Yahoo! Labs, Bangalore, India
mdixit@ucsd.edu, nikhil.rasiwasia@gmail.com, nuno@ece.ucsd.edu

## Abstract

*An extension of the latent Dirichlet allocation (LDA), denoted* class-specific-simplex *LDA (css-LDA), is proposed for image classification. An analysis of the supervised LDA models currently used for this task shows that the impact of class information on the topics discovered by these models is very weak in general. This implies that the discovered topics are driven by general image regularities, rather than the semantic regularities of interest for classification. To address this, we introduce a model that induces supervision in topic discovery, while retaining the original flexibility of LDA to account for unanticipated structures of interest. The proposed css-LDA is an LDA model with class supervision at the level of image features. In css-LDA topics are discovered per class, i.e. a single set of topics shared across classes is replaced by multiple class-specific topic sets. This model can be used for generative classification using the Bayes decision rule or even extended to discriminative classification with support vector machines (SVMs). A css-LDA model can endow an image with a vector of class and topic specific count statistics that are similar to the Bag-of-words (BoW) histogram. SVM-based discriminants can be learned for classes in the space of these histograms. The effectiveness of css-LDA model in both generative and discriminative classification frameworks is demonstrated through an extensive experimental evaluation, involving multiple benchmark datasets, where it is shown to outperform all existing LDA based image classification approaches.*

## 1. Introduction

*Bag-of-visual-words* (BoW) representation is quite popular in the image classification literature [12, 21]. Under BoW, the space of local descriptors are vector quantized with a set of representative points, known as "visual-words". An image is then summarized as a histogram of "visual word" co-occurrence [6]. For classification using image BoWs, the simplest architecture is equivalent to the *naive Bayes* approach to text classification [16]. It assumes that image *words* are sampled independently from the BoW model given the class, and relies on the Bayes decision rule for image classification. We refer to this as the *flat* model, due to its lack of hierarchical word groupings. Although capable of identifying sets of words discriminative for the classes of interest, it does not explicitly model the inter- and intra-class structure of word distributions. To facilitate the *discovery* of this structure, various models have been recently ported from the text to the vision literature. Popular examples include hierarchical *topic models*, such as latent Dirichlet allocation (LDA) [2] and probabilistic latent semantic analysis (pLSA) [9]. Under these, each document (or image) is represented as a finite mixture over an intermediate set of topics, which are expected to summarize the document semantics.

Since LDA and pLSA topics are discovered in an *unsupervised* fashion, these models have limited use for classification. Several LDA extensions have been proposed to address this limitation, in both the text and vision literatures. One popular extension is to apply a classifier, such as an SVM, to the topic representation [2, 3, 14]. We refer to these as discriminant extensions, and the combination of SVM with LDA topic vectors as SVM-LDA. Such extensions are hampered by the inability of *unsupervised* LDA to latch onto semantic regularities of interest for classification [1, 22]. For example, it has been noted in the text literature [1] that, given a collection of movie reviews, LDA might discover, as topics, movie properties, e.g. genres, which are not central to the classification task, e.g. prediction of movie ratings. A second approach is to incorporate a class label variable in the generative model [8, 1, 19, 11, 22, 13]. These are denoted generative extensions. Two popular members of this family are the model of [8], here referred to as classLDA (cLDA), and the model of [19], commonly known as supervisedLDA (sLDA). The latter was first proposed for supervised text prediction in [1]. Another popular generative extension is to directly equate the topics with the class labels themselves establishing a one-to-one mapping with between topics and class labels, e.g. labeled LDA [15], semiLDA [20]. We

*-indicates equal contribution

refer to such approaches as *topic-supervised* approaches.

In this work, we focus on generative extensions of LDA for image classification. We start by showing that even the most popular supervised extensions of LDA, such as cLDA and sLDA, are unlikely to capture class semantics. Theoretical analysis shows that the impact of class information on the topics discovered by cLDA and sLDA is very weak in general, and vanishes for large samples. Experiments demonstrate that the classification accuracies of cLDA and sLDA are not superior to those of unsupervised topic discovery. Next we extend the idea of topic-supervision to cLDA and sLDA models. Topic-supervision establishes a much stronger correlation between the topics and the class labels, nevertheless they are unable to outperform the simple flat model. In fact, we show that topic supervised models are fundamentally not different from the flat model.

To combine the *labeling strength* of topic-supervision with the *flexibility* of topic-discovery of LDA, we propose a novel classification architecture, denoted *class-specific simplex LDA* (css-LDA). Inspired by the flat model, css-LDA differs from the existing LDA extensions in that supervision is introduced directly at the level of image features. This induces the discovery of class-specific topic simplices and, consequently, *class-specific topic distributions*, enabling a much richer modeling of intra-class structure without compromising discrimination ability. Generative classification with the proposed model is performed using the Bayes decision rule. Experiments show that the css-LDA based classifier outperforms all the existing extensions of LDA and the flat model. We also extend the css-LDA model to a discriminative classification framework where it is used to infer a high dimensional vector of statistics per image. This image representation can be described as a set of topic specific word counts, where topics are informed by class labels. In the absence of topic structure and supervision, this vector reduces to the standard BoW histogram [6, 17]. When used with a discriminant SVM, the css-LDA based image representation is shown to produce significantly better results than the BoW histogram and as well as a topic-specific histogram derived from an unsupervised LDA model. Experimental evaluation is performed on five benchmark scene classification datasets[2].

## 2. Models of Image Classification

We start by reviewing LDA and its various extensions for classification. Images are represented as collections of visual words, $\mathcal{I} = \{w_1, \ldots, w_N\}, w_n \in \mathcal{V}$, where $\mathcal{V}$ is the 'codebook' of visual words. Each image in a *dataset*, $\mathcal{D} = \{\mathcal{I}_1, \ldots, \mathcal{I}_D\}$, is labeled with a class $y$, drawn from a random variable $Y$ with values in $\mathcal{Y} = \{1, \ldots, C\}$. This is the set of classes that define the image classification problem, making $\mathcal{D} = \{(\mathcal{I}_1, y_1), \ldots, (\mathcal{I}_D, y_D)\}$.

A query image $\mathcal{I}_q$ is classified with the minimum probability of error criterion, where the optimal decision rule is to assign $\mathcal{I}_q$ to the class of maximum posterior probability, i.e.

$$y^* = \arg\max_y P_{Y|W}(y|\mathcal{I}_q). \tag{1}$$

We next review some popular models.

### 2.1. Flat Model

Figure 1(a) presents the graphical form of the flat model. Visual words $w_n$ are sampled independently conditioned on the class label. The class prior $P_Y()$ and class-conditional distribution $P_{W|Y}()$ are chosen to be categorical distributions over $\mathcal{Y}$ and $\mathcal{V}$ respectively, The parameters $\Lambda^{flat}_{1:C}$ can be learned by maximum likelihood estimation as [6],

$$\Lambda^{flat}_{yw} = \frac{\sum_d \sum_n \delta(y^d, y)\delta(w^d_n, w)}{\sum_v \sum_d \sum_n \delta(y^d, y)\delta(w^d_n, v)}. \tag{2}$$

where $d$ indexes the training images, and $\delta(x, y)$ is the Kronecker delta function.

### 2.2. Unsupervised LDA Model

LDA is the unsupervised generative model shown in Figure 1(b). $P_\Pi()$ and $P_{W|Z}()$ are the prior and topic-conditional distributions respectively. $P_\Pi()$ is a Dirichlet distribution on $\mathcal{T}$ with parameter $\boldsymbol{\alpha}$, and $P_{W|Z}()$ a categorical distribution on $\mathcal{V}$ with parameters $\Lambda_{1:K}$. The model parameters are learned with a Variational Expectation Maximization (EM) algorithm [2]. Note that in its original formulation, LDA does not incorporate class information and cannot be used for classification.

### 2.3. Class LDA (cLDA)

ClassLDA (cLDA) was introduced in [8] for image classification. In this model, shown in Figure 1(c), a class variable $Y$ is introduced as the parent of the topic prior $\Pi$. In this way, each class defines a prior distribution in topic space $P_{\Pi|Y}(\boldsymbol{\pi}|y; \boldsymbol{\alpha}_y)$, conditioned on which the topic probability vector $\boldsymbol{\pi}$ is sampled. A query image $\mathcal{I}_q$ is classified with (1), using variational inference to approximate the posterior $P_{Y|W}(y|\mathcal{I}_q)$ [8].

### 2.4. Supervised LDA (sLDA)

sLDA was proposed in [1]. As shown in Figure 1(d), the class variable $Y$ is conditioned by topics $Z$. In its full generality, sLDA uses a generalized linear model of $Y$, which can be either discrete or continuous. [19] applied this generic framework to the task of image classification, where $Y$ takes on discrete responses, by making use of the softmax activation function. In this work, sLDA refers to the formulation of [19], since this was the one previously used for image classification. The class labels are sampled from $P_{Y|Z}()$ which is a softmax activation function with parameter $\boldsymbol{\zeta}_c \in \mathbb{R}^K$. Variational inference is used to learn all model parameters and to approximate the posterior $P_{Y|W}(y|\mathcal{I}_q)$, used in (1) for classifying an image $\mathcal{I}_q$ [19].

[2] Details regarding the datasets and our experimental protocol are provided in the supplement sec. IV

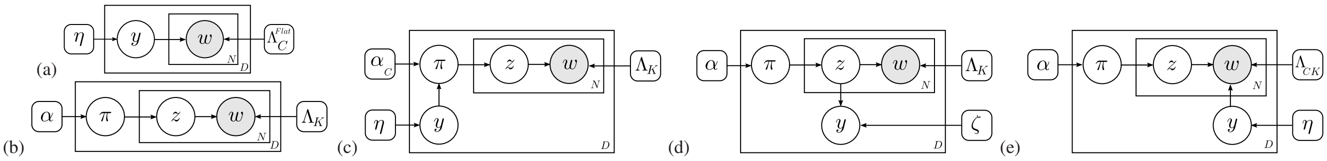
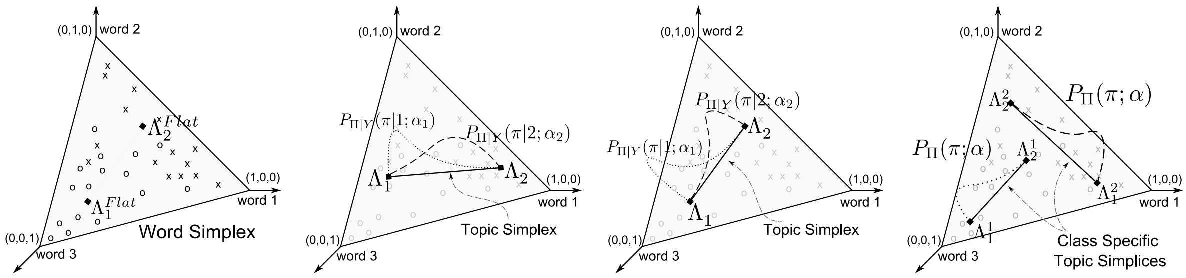


Figure 1. Graphical models for (a) flat model. (b) LDA and ts-LDA. (c) cLDA and ts-cLDA. (d) sLDA and ts-sLDA (e) css-LDA. All models use the plate notation of [4], with parameters shown in rounded squares.




(a) flat (b) cLDA (c) ts-cLDA (d) css-LDA

Figure 2. Representation of various models on a three word simplex. a) flat model. b) cLDA model with two topics. The line segment depicts a one-dimensional topic simplex. c) ts-cLDA model. Topic-conditional word distributions are learned with supervision and aligned with the class-conditional distributions of the flat model. d) css-LDA model. Each class defines its own topic simplex.

## 2.5. Topic-Supervised Approaches

Another popular approach to introduce supervision in LDA, is to equate topics directly to the class labels. The resulting extension is denoted as *topic supervised LDA* (ts-LDA) [20, 15]. The graphical model of the topic supervised extension of any LDA model is *exactly* the same as that of the model without topic supervision. The only, subtle yet significant, difference is that the topics are no longer discovered, but specified. This makes the topic-conditional distributions identical to the class-conditional distributions of the flat model. Topic-supervision for LDA model, was proposed in [20, 15]. In this work we also introduce the topic-supervised versions of cLDA and sLDA viz. ts-cLDA and ts-sLDA respectively.

## 2.6. Geometric Interpretation

The models discussed above have an elegant geometric interpretation [2, 18]. Associated with a vocabulary of $|\mathcal{V}|$ words, there is a $|\mathcal{V}|$ dimensional space, where each axis represents the occurrence of a particular word. A $|\mathcal{V}|-1$-simplex in this space, here referred to as *word simplex*, represents all probability distributions over words. Each image (when represented as a word histogram) is a point on this space. Figure 2(a) illustrates the two dimensional simplex of distributions over three words. Also shown are sample images from two classes, "o" from class-1 and "x" from class-2, and a schematic of the flat model. Under this model, each class is modeled by a class-conditional word distribution, i.e. a point on the word simplex. In Figure 2(a), $\Lambda_1^{flat}$ and $\Lambda_2^{flat}$ are the distributions of class-1 and class-2 respectively.

Figure 2(b) shows a schematic of cLDA with two topics. Each topic in an LDA model defines a probability distribution over words and is represented as a point on the word simplex. Since topic probabilities are mixing probabilities for word distributions, a set of $K$ topics defines a $K-1$ simplex in the word simplex, here denoted the *topic simplex*. If the number of topics $K$ is strictly smaller than the number of words $|\mathcal{V}|$, the topic simplex is a low-dimensional subsimplex of the word simplex. The projection of images on the topic simplex can be thought of as dimensionality reduction. In Figure 2(b), the two topics are represented by $\Lambda_1$ and $\Lambda_2$, and span a one-dimensional simplex, shown as a connecting line segment. In cLDA, each class defines a distribution (parameterized by $\boldsymbol{\alpha}_y$) on the topic simplex. The distributions of class-1 and class-2 are depicted in the figure as dotted and dashed lines, respectively. Similar to cLDA, sLDA can be represented on the topic simplex, where each class defines a softmax function[3].

Figure 2(c) shows the schematic of ts-cLDA for a two class problem on a three word simplex. As with cLDA, Figure 2(b), $\Lambda_1$ and $\Lambda_2$ are topic-distributions. There is, however, a significant difference. While the topic distributions of cLDA, learned by topic discovery, can be positioned anywhere on the word simplex, those of ts-cLDA are specified, and identical to the class-conditional distributions of the flat model.

[3]Strictly speaking, the softmax function is defined on the average of the sampled topic assignment labels $\bar{\mathbf{z}}$. However, when the number of features $N$ is sufficiently large, $\bar{\mathbf{z}}$ is proportional to the topic distribution $\boldsymbol{\pi}$. Thus, the softmax function can be thought of as defined on the topic simplex.

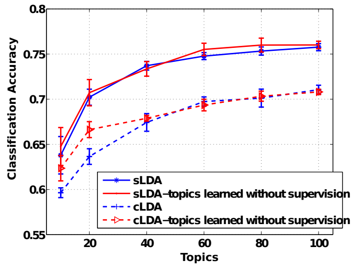


Figure 3. Classification accuracy as function of the number of topics for sLDA and cLDA, using topics learned with and without class influence and codebooks of size 1024 on N13. Similar behavior was observed for codebooks of different sizes and different datasets.

## 3. Limitations of Existing Models

In this section we present theoretical and experimental evidence that, contrary to popular belief, topics discovered by sLDA and cLDA are not more suitable for discrimination than those of standard LDA. We start by analyzing the variational EM algorithms for the two.

In both sLDA and cLDA the parameters $\Lambda_{1:K}$ of the topic distributions are obtained via the variational M-step as $\Lambda_{kv} \propto \sum_d \sum_n \delta(w_n^d, v)\phi_{nk}^d$, where $d$ indexes the images, $\sum_v \Lambda_{kv} = 1$, $\delta()$ is a Kronecker delta function and $\phi_{nk}$ is the parameter of the variational distribution $q(z)$. This parameter is computed in the E-step.

For cLDA:
$$\gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_{y^d k} \quad (3)$$
$$\phi_{nk}^{d*} \propto \Lambda_{k w_n^d} \exp\left[\psi(\gamma_k^d)\right] \quad (4)$$
For sLDA:
$$\gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_k \quad (5)$$
$$\phi_{nk}^{d*} \propto \Lambda_{k w_n^d} \exp\Bigg[\psi(\gamma_k^d) + \frac{\zeta_{y^d k}}{N} - \frac{\sum_c \exp\frac{\zeta_{ck}}{N} \prod_{m\neq n} \sum_j \phi_{mj}^d \exp\frac{\zeta_{cj}}{N}}{\sum_c \prod_m \sum_j \phi_{mj}^d \exp\frac{\zeta_{cj}}{N}}\Bigg] \quad (6)$$

where $\boldsymbol{\gamma}$ is the parameter of the variational distribution $q(\boldsymbol{\pi})$ (see [2] for the details of variational inference in LDA). The important point is that the class label $y^d$ only influences the topic distributions through (3) for cLDA (where $\boldsymbol{\alpha}_{y^d}$ is used to compute the parameter $\boldsymbol{\gamma}^d$) and (6) for sLDA (where the variational parameter $\phi_{nk}^d$ depends on the class label $y^d$ through $\zeta_{y^d k}/N$).

We next consider the case of cLDA. Given that $q(\boldsymbol{\pi})$ is a posterior Dirichlet distribution (and omitting the dependence on $d$ for simplicity), the estimate of $\gamma_k$ has two components: $\hat{l}_k = \sum_{n=1}^N \phi_{nk}$, which acts as a vector of counts, and $\alpha_{yk}$ which is the parameter from the prior distribution. As the number of visual words $N$ increases, the amplitude of the count vector, $\hat{\mathbf{l}}$, increases proportionally, while the prior $\boldsymbol{\alpha}_y$ remains constant. Hence, for a sufficiently large sample size $N$, the prior $\boldsymbol{\alpha}_y$ has a very weak influence on the estimate of $\boldsymbol{\gamma}$. This is a hallmark of Bayesian parameter estimation, where the prior only has impact on the posterior estimates for small sample sizes. It follows that the connection between class label $Y$ and the learned topics $\Gamma_k$ is extremely weak. This is not a fallacy of the variational approximation. In cLDA (Figure 1(b)), the class label distribution is simply a prior for the remaining random variables. This prior is easily overwhelmed by the evidence collected at the feature-level, whenever the sample is large. A similar effect holds for sLDA, where the only dependence of the parameter estimates on the class label is through the term $\zeta_{y^d k}/N$. This clearly diminishes as the sample size $N$ increases[4]. In summary, topics learned with either cLDA or sLDA are very unlikely to be informative of semantic regularities of interest for classification, and much more likely to capture generic regularities, common to all classes.

To confirm these observations, we performed experiments with topics learned under two approaches. The first used the original learning equations, i.e. (3) and (4) for cLDA and (5) and (6) for sLDA. In the second we severed all connections with the class label variable *during topic learning*, by reducing the variational E-step (of both cLDA and sLDA) to,

$$\gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha, \quad \phi_{nk}^{d*} \propto \Lambda_{k w_n^d} \exp\left[\psi(\gamma_k^d)\right] \quad (7)$$

with $\alpha = 1$. This guarantees that the topic-conditional distributions are learned without any class influence. The remaining parameters ($\boldsymbol{\alpha}_y$ for cLDA, $\boldsymbol{\zeta}_y$ for sLDA) are still learned using the original equations. The rationale for these experiments is that, if supervision makes any difference, models learned with the original algorithms should perform better.

Figure 3 shows the image classification performance of cLDA and sLDA, under the two learning approaches on the N13 dataset [8] (sec. IV in supplement). The plots were obtained with a 1024 words codebook, and between 10 and 100 topics. Clearly, the classification performance of the original models *is not* superior to that of the ones learned without class supervision. The sLDA model has almost identical performance under the two approaches. Similar effects were observed in experiments with codebooks of different size and on different datasets. These results show that the performance of cLDA and sLDA is similar to that of topic learning without class supervision. In both cases, the class variable has very weak impact on the learning of topic distributions.

### 3.1. Limitations of Topic-Supervised Models

In the previous section, we have seen that models such as sLDA or cLDA *effectively* learn topics without supervision. The simplest solution to address the lack of correlation between class labels and the topics, is to *force* topics to reflect the semantic regularities of interest as is done in

[4]This discussion refers to the sLDA formulation of [19], which was proposed specifically for image classification.

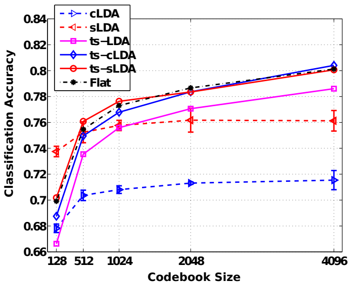



Figure 4. Classification accuracy vs. codebook size for ts-sLDA, ts-cLDA, sLDA, cLDA, and flat model on N13. For ts-sLDA and ts-cLDA the number of topics is equal to the number of classes. For sLDA and cLDA, results are presented for the number of topics of best performance.

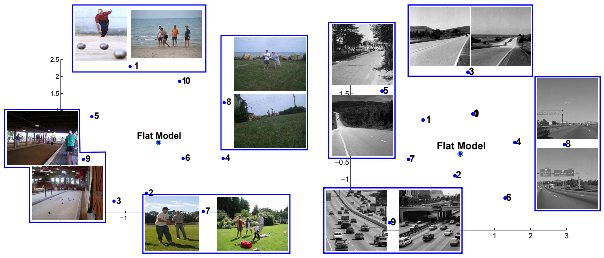



Figure 5. Two-dimensional embedding of the topic vectors discovered by css-LDA (marked #1 - #10), and class-conditional distribution of flat model (marked flat model), for left) "Bocce"(S8) and right) "Highway"(N13) classes. Also shown are the nearest neighbor images of sample topic conditional distributions.

topic-supervised(ts-) models. Figure 4 presents classification results of ts-LDA, ts-cLDA and ts-sLDA, as a function of codebook size, under the experimental conditions of Figure 3. Also shown are the accuracies of cLDA, sLDA, and the flat model. It is clear that, although outperforming their unsupervised counterparts, topic-supervised models cannot beat the flat model. This is troubling, since such modeling increases the complexity of both LDA learning and inference, which are certainly larger than those of the flat model. It places in question the usefulness of the whole LDA approach to image classification. This problem has simply been ignored in the literature, where comparisons of LDA based approaches with the flat model are usually not presented.

## 4. Class Specific Simplex Latent Dirichlet Allocation (css-LDA)

To overcome the limitation of existing LDA based image classification models, in this section we introduce a new LDA model for image classification, denoted class-specific simplex LDA.

### 4.1. Motivation

The inability of the LDA variants to outperform the flat model is perhaps best understood by returning to Figure 2. Note that both cLDA and ts-cLDA map images from a high dimensional word simplex to a low dimensional topic simplex, which is common to all classes. This restricts the scope of the class models, which are simple Dirichlet distributions over the topic simplex. Similar pictures hold for sLDA and ts-sLDA, where the classes define a softmax function in the simplex. In fact, even SVM-LDA learns an SVM classifier on this space. Since the topic simplex is common, and low dimensional, too few degrees of freedom are available to characterize intra-class structure, preventing a very detailed discrimination of the different classes. In fact, the main conclusion of the previous sections is that the bulk of the modeling power of LDA lies on the selection of the topic simplex, and not on the modeling of the data distribution in it. Since to capture the semantic regularities of the data, the simplex has to be aligned with the class labels — as is done under topic-supervision — there is little room to outperform the flat model.

This limitation is common to any model that constrains the class-conditional distributions to lie on a common topic simplex. This is the case whenever the class label $Y$ is connected to either the prior $\Pi$ or topic $Z$ variables, as in the graphical models of Figure 1. Since the topic simplex is smaller than the word simplex, it has limited ability to simultaneously model rich intra-class structure and keep the classes separated. For this, it is necessary that the class label $Y$ affects the word distributions *directly,* freeing these to distribute themselves across the word simplex in the most discriminant manner. This implies that $Y$ must be connected to the word variable $W$, as in the flat model. The result is the graphical model of Figure 1(e), which turns out to have a number of other properties of interest.

The first follows from the fact that it makes the topic conditional distributions dependent on the class. Returning to Figure 2, this implies that the vertices of the topic simplex are class-dependent, as shown in (d). Note that there are two one-dimensional topic simplices, one for each class defined by the parameters $\Lambda_1^y$ and $\Lambda_2^y$, $y \in \{1, 2\}$. The dotted and dashed lines denote the prior distribution on the topic simplices, which is controlled by the $\boldsymbol{\alpha}$ parameter. Hence, each class is endowed with its own topic simplex justifying the denomination of the model as *class-specific simplex LDA.*

Second, css-LDA extends both the flat and the LDA model, simultaneously addressing the main limitations of these two models. With respect to LDA, because there are *multiple topic simplices*, the class-conditional distributions can have little overlap in word-simplex even when topic simplices are low dimensional. Since the simplex is different from those of other classes, this does not compromise discrimination. On the other hand, because a much larger set of topic distributions is now possible per class, the model has much greater ability to model intra-class structure.

With respect to the flat model, css-LDA inherits the advantages of LDA over bag-of-words. Consider the collection of images from the class "Bocce" of the S8 dataset shown in the left side of Figure 5. Note that the game of Bocce can be played indoors or outdoors, on beach or on grass, on an overcast day or a sunny day, etc. Each of these conditions leads to drastically different scene appearances and thus a great diversity of word distributions. Under the flat model, the "Bocce" class is modeled by a single point in the word simplex, the average of all these distributions, as shown in Figure 2 (a). This is usually insufficient to capture the richness of each class. Rather than this, css-LDA devotes to each class a topic simplex, as shown in Figure 2 (d). This increases the expressive power of the model, because there are now many topic-conditional word distributions per class. In the example of Figure 2, while the flat model approximates all the images of each class by a point in word simplex, css-LDA relies on a line segment. In higher dimensions the difference can be much more substantial, since each topic simplex is a subspace of dimension $K - 1$ ($K$ the number of topics), while the approximation of the flat model is always a point. Thus css-LDA can account for much more complex class structure than the flat counterpart.

### 4.2. The css-LDA model

The graphical model of css-LDA model is shown in Figure 2(e). Similar to the earlier LDA extensions, $P_Y()$ is a categorical distribution over $\mathcal{Y}$ with parameter $\boldsymbol{\eta}$, $P_\Pi()$ a Dirichlet distribution on the topic simplex with parameter $\boldsymbol{\alpha}$, $P_{Z|\Pi}()$ a categorical distribution over $\mathcal{T}$ with parameter $\boldsymbol{\pi}$ and $P_{W|Z,Y}()$ a categorical distribution over $\mathcal{V}$ with a class dependent parameter $\Lambda_z^y$.

Like previous models, learning and inference are intractable. Given an image $\mathcal{I} = \{w_1, \ldots, w_N\}, w_n \in \mathcal{V}$, inference consists of computing the posterior distribution

$$P(y, \boldsymbol{\pi}, z_{1:N}|\mathcal{I}) = P(\boldsymbol{\pi}, z_{1:N}|\mathcal{I}, y)P(y|\mathcal{I}) \quad (8)$$

$$\text{where,} \quad P_{Y|W}(y|\mathcal{I}) = \frac{P_{Y,W}(y, \mathcal{I})}{\sum_c P_{Y,W}(c, \mathcal{I})}. \quad (9)$$

Both $P_{Y,W}(y, \mathcal{I})$ and $P_{\Pi,Z|W,Y}(\boldsymbol{\pi}, z_{1:N}|\mathcal{I}, y)$ are intractable and approximated using variational methods. The posterior $P_{\Pi,Z|W,Y}(\boldsymbol{\pi}, z_{1:N}|\mathcal{I}, y)$ is approximated by the variational distribution $q(\boldsymbol{\pi}, z_{1:N}) = q(\boldsymbol{\pi}; \boldsymbol{\gamma}) \prod_n q(z_n; \boldsymbol{\phi}_n)$. The marginal likelihood $P_{Y,W}(y, \mathcal{I})$ is approximated by maximizing the evidence lower bound $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\eta}, \boldsymbol{\alpha}, \Lambda)$ for different values of $y \in \{1, \ldots, C\}$, i.e. $P_{W,Y}(\mathcal{I}, y) \sim \max_{\boldsymbol{\gamma},\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\eta}, \boldsymbol{\alpha}, \Lambda)$ where $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\eta}, \boldsymbol{\alpha}, \Lambda) = E_q[\log P(y, \boldsymbol{\pi}, z_{1:N}, w_{1:N})] - E_q[\log q(\boldsymbol{\pi}, z_{1:N})]$. Solving for $\mathcal{L}$ for a given $y$, results in updates similar to the standard LDA inference equations (sec. I in supplement),

$$\gamma_k^* = \sum_n \phi_{nk} + \alpha_k, \qquad \phi_{nk}^* \propto \Lambda_{kw_n}^y \exp\left[\psi(\gamma_k)\right] \quad (10)$$

Note that for css-LDA, where each class is associated with a separate topic simplex, (10) differs from standard LDA in that the $\Lambda$ parameters are class specific.

Model learning involves estimating the parameters $(\boldsymbol{\eta}, \boldsymbol{\alpha}, \Lambda_{1:K}^{1:C})$ by maximizing the log likelihood, $l = \log P_W(\mathcal{D})$, of a training image data $\mathcal{D}$. This is achieved using a variational EM algorithm [2] that follows the update rules of (10) in the variational E-step (sec. III in supplement). In our experiments, we found that the performance of css-LDA was not very sensitive to the choice of the parameter $\boldsymbol{\alpha}$. Therefore, instead of learning $\boldsymbol{\alpha}$, it was always set to a fixed value. In this setting, a css-LDA is equivalent to learning a separate LDA model for each class, with a fixed Dirichlet prior.

### 4.3. The css-LDA image representation

In the generative classification framework, all the models discussed so far, can be used directly with the Bayes decision rule. For discriminative classification, these models help to endow images with histograms of their visual words. Class decision boundaries are then, learned as discriminants in the space of these histograms. If an image BoW with N words, is modeled as a flat categorical distribution with parameters $\Lambda_v$ where $v \in \{1 \ldots |\mathcal{V}|\}$, and $\Lambda_v = \exp(\theta_v)/\sum_{v'} \exp(\theta_{v'})$, the gradient of its log-likelihood, can be shown to produce a histogram of visual words [5].

$$\frac{\partial \mathcal{L}}{\partial \theta_v} \propto \sum_n \delta(w_n, v) \quad (11)$$

We denote this image histogram as a BoW-vec. In the vision literature, a BoW-vec has been widely used along with SVM classifiers to produce impressive results [6, 12]. Similar image representations can be obtained by modeling images using LDA based models. Under LDA, however, the gradient of image log-likelihood is approximated by the gradient of its lower bound obtained by variational inference. For the basic LDA model of section 2.2, the gradients with respect to its parameters $\Lambda_{kv}$, produce a topic specific visual-word histogram expressed as,

$$\frac{\partial \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \Lambda)}{\partial \Lambda_{kv}} \propto \sum_n \phi_{nk} \delta(w_n, v) \quad (12)$$

This vector, referred here as the LDA-vec, is higher dimensional compared to the BoW-vec, because of the larger parameter space of LDA compared to the flat model. The css-LDA model, proposed in this work, models image words with distinct class specific simplices of topics as shown in fig 2(d). The gradients of its evidence lower bound, therefore, produce an even larger histogram with class and topic specific word counts given an image.

$$\frac{\partial \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\eta}, \boldsymbol{\alpha}, \Lambda)}{\partial \Lambda_{kv}^y} \propto \sum_n \phi_{nk}^y \delta(w_n, v) \quad (13)$$

The resulting $(C * T * |\mathcal{V}|)$ dimensional css-LDA-vec is much larger than the $(T * |\mathcal{V}|)$ dimensional LDA-vec and the $|\mathcal{V}|$ dimensional BoW-vec histograms. In the degenerate case with a single class and a single topic, $C = 1$ and $\phi_{n1} =$

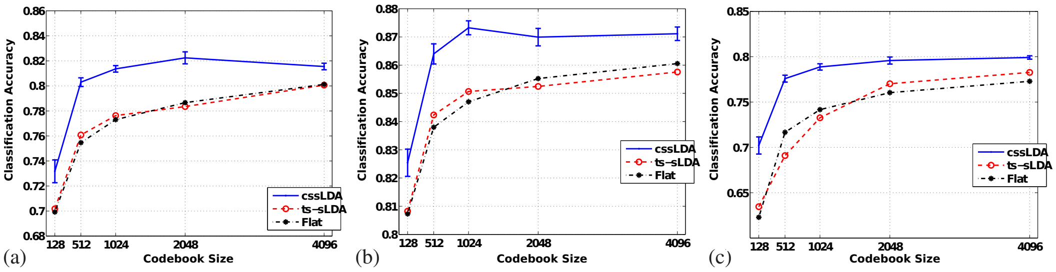



Figure 6. Classification accuracy of css-LDA, ts-sLDA and the flat model, as a function of codebook size on (a) N13, (b) N8 and (c) S8. The reported css-LDA performance is the best across number of topics, while for ts-sLDA the number of topics is equal to the number of classes.

1, the LDA counts of (13) and (12), reduce to the BoW counts of (11). In our experiments, we use these histograms to train linear SVMs for classification.

## 5. Results

Several experiments were performed to evaluate css-LDA, on standard scene classification datasets. These include the 8 class (N8), 13 class (N13) and 15 class (N15) datasets previously used in [8, 3, 14, 12], the 8 class sports dataset (S8) of [19] and a 50 class dataset (C50) constructed from the Coral image collection used in [7]. The experimental setup is discussed in sec. IV in the supplement.

### 5.1. Class Specific Topic Discovery in css-LDA

We start with a set of experiments that provide insight on the topics discovered by css-LDA. Figure 5 presents a visualization of the topic-conditional distributions $\Lambda_z^y$ (marked #1 to #10) discovered for classes "Bocce" (S8, left) and "Highway" (N13, right), using 10 topics per class. Also shown is the class conditional distribution $\Lambda_y^{flat}$ (marked flat model) of the flat model. The visualization was produced by a 2D embedding of the word simplex, using non-metric multidimensional scaling [10] from the matrix of KL divergences between topic- and class-conditional distributions. Note how, for both classes, the flat model is very close to the average of the topic-conditional distributions. This shows that, on average, topics discovered by css-LDA represent the class conditional distribution of the flat model. In fact, the KL divergence between the average of the topic conditional distributions of css-LDA and the class conditional distribution of the flat model is very close to zero ($0.013 \pm 0.019$ for N13, $0.014 \pm 0.008$ for S8). Also shown in the figure, for some sample topics, are the two images closest to the topic conditional distribution. Note that the topics discovered by css-LDA capture the visual diversity of each class. For example, "Bocce" topics #9, #7, #8 and #1 capture the diversity of environments on which sport can be played: indoors, sunny-outdoor, overcast-outdoor, and beach. These variations are averaged out by the flat model, where each class is, in effect, modeled by a single topic.

### 5.2. Generative Classification Results

We have previously reported that all the known LDA models are outperformed by their topic supervised (ts-) extensions. Figure 6, however, indicates that the performance of ts-sLDA (the best performer amongst the (ts-) models) is very close to the flat model for datasets N13, N8 and S8. This is in stark contrast to css-LDA, which has a clearly better performance than the two, across datasets and codebook sizes. Since css-LDA is an extension of the flat model, this gain can be attributed to its topic-discovery mechanism.

Table 1 summarizes the best classification accuracy achieved by all methods considered in this work, plus SVM-LDA [2, 3, 14], on all datasets. Note that css-LDA outperforms all existing generative models for image classification, and a discriminative classifier, SVM-LDA, on all five datasets, with a classification accuracy of 76.62% on N15, 81.03% on N13, 87.97% on N8, 80.37% on S8 and 46.04% on C50. On average it has a relative gain of 3.0% over the flat model, 3.5% over ts-sLDA, 4.9% over ts-cLDA, 6.7% over ts-LDA, 8.5% over sLDA, 17.2% over cLDA and 4.0% over SVM-LDA.

### 5.3. Discriminative Classification Results

To evaluate the discriminative classification performance of the BoW-vec, the LDA-vec and the css-LDA-vec representations of (11), (12) and (13), we use them with a linear SVM. Generative models that produce these vectors, are learned using visual vocabulary of size 1024 and 30 LDA topics. All three representations are $L_1$ normalized and square-rooted. The square root operation is known to induce a Hellinger kernel over histogram, on taking a dot-product.

Results reported in Table 1 show that the css-LDA-vec significantly outperforms both BoW-vec and the LDA-vec representations with average accuracies of 78.19% on N15, 81.44% on N13, 88.10% on N8, 81.67% on S8 and 53.4% on C50[5]. Although, the LDA-vec is higher dimensional than the BoW-vec, it is not clearly better in terms of performance. This indicates that the gains achieved by the

[5]Similar experiments were performed using an Intersection kernel SVM (IKSVM). The gains of css-LDA-vec, over LDA-vec and BoW-vec, with IKSVM, were almost the same as those reported in Table 1. Due to lack of space, however, these results could not be included here.

Table 1. Generative and Discriminative Classification Results.

| model | Dataset | | | | |
|---|---|---|---|---|---|
| | **N15** | **N13** | **N8** | **S8** | **C50** |
| css-LDA | **76.62 ± 0.32** | **81.03 ± 0.74** | **87.97 ± 0.84** | **80.37 ± 1.36** | **46.04** |
| flat | 74.91 ± 0.38 | 79.60 ± 0.38 | 86.80 ± 0.51 | 77.87 ± 1.18 | 43.20 |
| ts-sLDA | 74.82 ± 0.68 | 79.70 ± 0.48 | 86.33 ± 0.69 | 78.37 ± 0.80 | 42.33 |
| ts-cLDA | 74.38 ± 0.78 | 78.92 ± 0.68 | 86.25 ± 1.23 | 77.43 ± 0.97 | 40.80 |
| ts-LDA | 72.60 ± 0.51 | 78.10 ± 0.31 | 85.53 ± 0.41 | 77.77 ± 1.02 | 39.20 |
| medLDA [22] | 72.08 ± 0.59 | 77.58 ± 0.58 | 85.16 ± 0.57 | 78.19 ± 1.05 | 41.89 |
| sLDA [19] | 70.87 ± 0.48 | 76.17 ± 0.92 | 84.95 ± 0.51 | 74.95 ± 1.03 | 39.22 |
| cLDA [8] | 65.50 ± 0.32 | 72.02 ± 0.58 | 81.30 ± 0.55 | 70.33 ± 0.86 | 34.33 |
| css-LDA-vec | **78.19 ± 0.71** | **81.44 ± 0.50** | **88.10 ± 0.60** | **81.67 ± 1.93** | **53.4** |
| LDA-vec | 76.68 ± 0.33 | 80.15 ± 0.63 | 87.20 ± 0.48 | 80.17 ± 1.07 | 49.4 |
| BoW-vec | 76.66 ± 0.71 | 80.24 ± 0.70 | 87.47 ± 0.50 | 79.34 ± 0.93 | 46.4 |
| LDA-SVM | 73.19 ± 0.51 | 78.45 ± 0.34 | 86.82 ± 0.93 | 76.32 ± 0.71 | 45.46 |

css-LDA-vec are due to the discriminative nature of the underlying model and not just the larger size of its parameter space.

## 6. Conclusion

In this work we demonstrated the weakness of the existing LDA models in modeling discriminative information necessary for image classification. To address this, we proposed a novel css-LDA, model which induces a topic-simplex per class, thus providing a greater flexibility of modeling discriminant information. As a generative classifier, the css-LDA model was shown to outperform all its known LDA counterparts. In the discriminative classification framework, the css-LDA based image histogram was shown superior to the alternative image representations based on flat model and the unsupervised LDA.


### Acknowledgement

This research was supported by NSF awards CCF-0830535 and IIS-1208522.


## References


[1] D. Blei and J. McAuliffe. Supervised topic models. *NIPS*, 20:121–128, 2008. 1, 2

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003. 1, 2, 3, 4, 6, 7

[3] A. Bosch, A. Zisserman, and X. Muoz. Scene classification using a hybrid generative/discriminative approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(4):712–727, 2008. 1, 7

[4] W. Buntine. Operations for learning with graphical models. *Arxiv preprint cs/9412102*, 1994. 3

[5] R. Cinbis, J. Verbeek, and C. Schmid. Image categorization using fisher kernels of non-iid image models. In *IEEE CVPR 2012*, 2012. 6

[6] G. Csurka, C. Dance, L. Fan, and C. Bray. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, 1:1–22, 2004. 1, 2, 6

[7] P. Duygulu, K. Barnard, N. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002. 7

[8] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *IEEE CVPR*, 2005. 1, 2, 4, 7, 8

[9] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999. 1

[10] J. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, 1964. 7

[11] S. Lacoste-Julien, F. Sha, and M. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. *NIPS*, 21, 2008. 1

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE CVPR*, 2006. 1, 6, 7

[13] D. Putthividhya, H. Attias, and S. Nagarajan. Supervised topic model for automatic image annotation. In *IEEE ICASSP*, 2010. 1

[14] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(9):1575–1589, 2007. 1, 7

[15] D. Ramage, D. Hall, R. Nallapati, and C. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256. Association for Computational Linguistics, 2009. 1, 3

[16] J. Rennie. *Improving Multi-class Text Classification with Naive Bayes*. PhD thesis, Massachusetts Institute of Technology, 2001. 1

[17] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *IEEE ICCV*, 2003. 2

[18] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007. 3

[19] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *IEEE CVPR*, 2009. 1, 2, 4, 7, 8

[20] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent Dirichlet allocation: A hierarchical model for human action recognition. *Human Motion Understanding, Modeling, Capture and Animation*, pages 240–254, 2007. 1, 3

[21] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *IEEE ICCV*, 2005. 1

[22] J. Zhu, A. Ahmed, and E. Xing. MedLDA: maximum margin supervised topic models for regression and classification. In *ICML*, 2009. 1, 8