Chapter 7

The Importance of Supervision

7.1 Introduction

The architecture proposed in Chapter 6 has several properties in common with the family of *theme* or *topic models,* [14, 58]. Topic models were introduced to facilitate the *discovery* of hidden structure in a corpus of data in the text processing literature. Popular examples include latent Dirichlet allocation (LDA) [14] and probabilistic latent semantic analysis (pLSA) [58]. In these models, each entry in a corpus is represented as a finite mixture over an intermediate set of *topics* discovered in an *unsupervised* fashion. However, in their original formulations, topic models do not incorporate supervised information and can not be directly employed for classification.

Several extensions of the LDA model have been proposed to address this limitation in both the text and vision literatures[1]. One popular extension is to apply a classifier, such as a SVM, to the topic representation learned by these models [14, 17, 114]. A second approach is to incorporate a class label variable in the generative model [77, 13, 167, 71, 180, 112]. These are denoted generative extensions. Two popular extensions in this family, for scene classification, are that of [77], here referred to as classLDA (cLDA), and [167], commonly known as supervisedLDA (sLDA). The latter was first proposed for supervised text prediction in [13]. Thus, like the representation of holistic context models, topic models for supervised tasks have two layers. Appearance features are used to compute topic probabilities (that correspond to the proposed SMNs), which are hierarchically propagated to a more abstract layer that computes class probabilities (correspondent to the proposed CMNs).

In this chapter, we discuss the generative extensions of the LDA model in context of the proposed holistic context models (see Chapter 6). We start by highlighting the similarities and differences between the cLDA model and the holistic context model. Although the Bayesian network for both these models are very similar, there are fundamental differences, the most important being the level of supervision. Existing generative extensions of LDA such as cLDA and sLDA

[1]Note that some of these models were discussed in Chapter 4, however for clarity of the presentation these models are reviewed again in this chapter

rely on unsupervised discovery of topic. This fundamentally restricts their efficacy for the task of visual recognition. This is shown by 1) a theoretical analysis of the learning algorithms, and 2) experimental evaluation on classification problems. Theoretically, it is shown 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 show that the classification accuracies of cLDA and sLDA *are not superior* to those of unsupervised topic discovery. Although the holistic context models are effective at addressing this limitation, they have a different learning and inference procedure, which prevent a systematic study of the benefits of supervision in these models. Infact, existing approaches rely on the bag-of-words representation whereas bag-of-features was the choice of image representation in holistic context model (see 2.1.1 for details). In this chapter, to test the benefits of supervision in LDA models, we propose a family of LDA models which we denote as *topic supervised (ts)*. Instead of relying on *discovered* topics, topic-supervised LDA *equates topics to the classes of interest* for scene classification, establishing a one-to-one mapping between topics and class labels. This *forces LDA to pursue semantic regularities in the data*.

Note that the only, subtle yet significant, difference between the existing generative extensions and the proposed topic supervised extensions, is that the topics are no longer *discovered*, but *specified*. Both these systems rely on the same image representation, that of bag-of-words, and the same learning/inference procedures (although as we shall see in 7.5.3, learning in topic supervised models is much more simplified). This enables us to attribute any difference in their performance, to the difference in the level of supervision. It is shown that, topic supervision significantly improves on the classification accuracy of existing supervised LDA extensions. This is demonstrated by the introduction of *topic supervised* versions of LDA, cLDA and sLDA, denoted *ts-LDA*, *ts-cLDA* and *ts-sLDA* respectively. In all cases, the performance of topic supervised models is superior to that of the corresponding LDA models learned without topic-supervision.

The chapter is organized as follows. Section 7.2 briefly reviews the literature on generative models for scene classification. Topic models, in particular cLDA

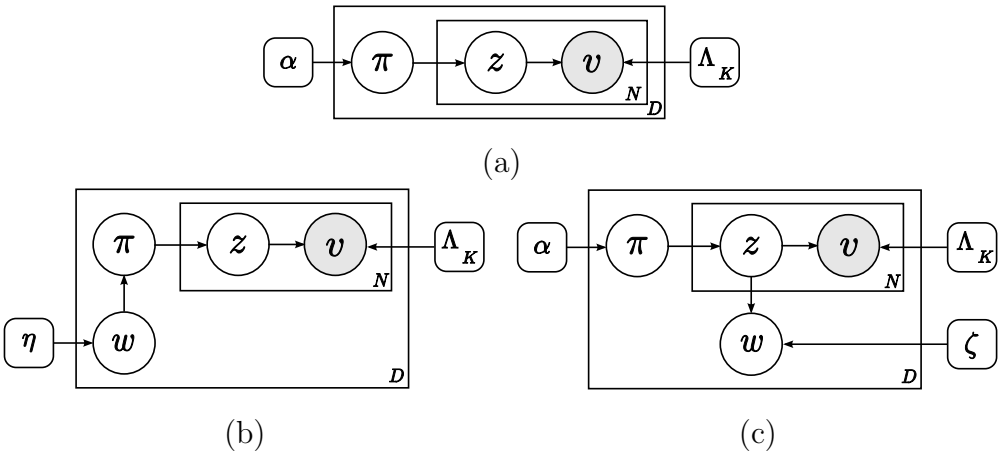


Figure 7.1: Graphical models for (a) LDA and ts-LDA. (b) cLDA and ts-cLDA. (c) sLDA and ts-sLDA. All models use the standard plate notation [19], with parameters shown in rounded squares.

model is compared to the holistic context models in Section 7.3. The limitations of existing models are highlighted in Section 7.4. Next, in Section 7.5 we introduce the topic-supervised model. An extensive experimental evaluation of the proposed frameworks is presented in Sections 7.5.4.

7.2 Topic Models

We start by reviewing LDA and its various generative extensions for classification.

7.2.1 LDA model

LDA is the generative model of Figure 7.1(a). Under it, images are sampled as follows.

for each image **do**
  sample $\boldsymbol{\pi} \sim P_{\Pi}(\boldsymbol{\pi}; \boldsymbol{\alpha})$.
  for $i \in \{1, \ldots, N\}$ **do**
    sample a topic, $z_i \sim P_{Z|\Pi}(z_i|\boldsymbol{\pi})$, $z_i \in \mathcal{L} = \{1, \ldots, K\}$, where $\mathcal{L}$ is the set of topics.

sample a visual word $v_i \sim P_{V|Z}(v_i|z_i;\Lambda_{z_i})$.

end for

end for

where $P_\Pi()$ and $P_{V|Z}()$ are the prior and topic-conditional distributions respectively. $P_\Pi()$ is a Dirichlet distribution on $\mathcal{L}$ with parameter $\boldsymbol{\alpha}$, and $P_{V|Z}()$ a categorical distribution on $\mathcal{V}$ with parameters $\Lambda_{1:K}$. Although the parameters of the model can be learned with the well known expectation maximization (EM) algorithm, the E-step yields an intractable inference problem. To address this, a wide range of approximate inference methods have been proposed [11], such as Laplace or variational approximations, sampling methods, etc. In this work, we adopt variational inference for all models where exact inference is intractable. Variational inference for the LDA model is briefly discussed in Appendix D[2]. In its original formulation, LDA does not incorporate class information and cannot be used for classification. We next discuss two models proposed to address this limitation.

7.2.2 Class LDA (cLDA)

ClassLDA (cLDA) was introduced in [77] for image classification. In this model, shown in Figure 7.1(b), a class variable W is introduced as the parent of the topic prior Π. In this way, each class defines a prior distribution in topic space, conditioned on which the topic probability vector $\boldsymbol{\pi}$ is sampled. Images are sampled as follows

for each image **do**

sample a class label $w \sim P_W(w;\boldsymbol{\eta})$, $w \in \mathcal{W}$

sample $\boldsymbol{\pi} \sim P_{\Pi|W}(\boldsymbol{\pi}|w;\boldsymbol{\alpha}_w)$.

for $i \in \{1,\ldots,N\}$ **do**

sample a topic, $z_i \sim P_{Z|\Pi}(z_i|\boldsymbol{\pi})$, $z_i \in \mathcal{L} = \{1,\ldots,K\}$.

sample a visual word $v_i \sim P_{V|Z}(v_i|z_i;\Lambda_{z_i})$

[2]Note that the variational inference procedure is detailed for the LDA model of Figure 2.5(b), which has notational differences with Figure 7.1(a), but the variational inference procedure is identical.

end for

end for

where, $\boldsymbol{\alpha}_w = \{\alpha_{w1}, \ldots, \alpha_{wK}\}$. Parameter learning for cLDA is similar to that of LDA [77] and detailed in Appendix E.

Given image $\mathcal{I}_q$, classification is performed by MPE decision rule, where the posterior $P_{W|V}(w|\mathcal{I}_q)$ can be approximated using a variational approximation [77].

7.2.3 Supervised LDA (sLDA)

The sLDA model was proposed in [13]. As shown in Figure 7.1(c), the class variable W is conditioned by the topics Z. The original formulation uses unconstrained real-valued response variables W and is not suitable for classification. An extension to discrete responses, using a softmax function, was introduced in [167]. An alternative extension to binary image annotation was proposed in [112], using a multi-variate Bernoulli variable for W. In [180], the max-margin principle is used to train sLDA, which is denoted maximum entropy discrimination LDA (medLDA). In this work, sLDA refers to the formulation of [167], since this was the one previously used for scene classification. Images are sampled as follows

for each image **do**

sample $\boldsymbol{\pi} \sim P_{\Pi}(\boldsymbol{\pi}; \boldsymbol{\alpha})$.

for $i \in \{1, \ldots, N\}$ **do**

sample a topic, $z_i \sim P_{Z|\Pi}(z_i|\boldsymbol{\pi})$, $z_i \in \mathcal{L} = \{1, \ldots, K\}$

sample a visual word $v_i \sim P_{V|Z}(v_i|z_i; \Lambda_{z_i})$.

end for

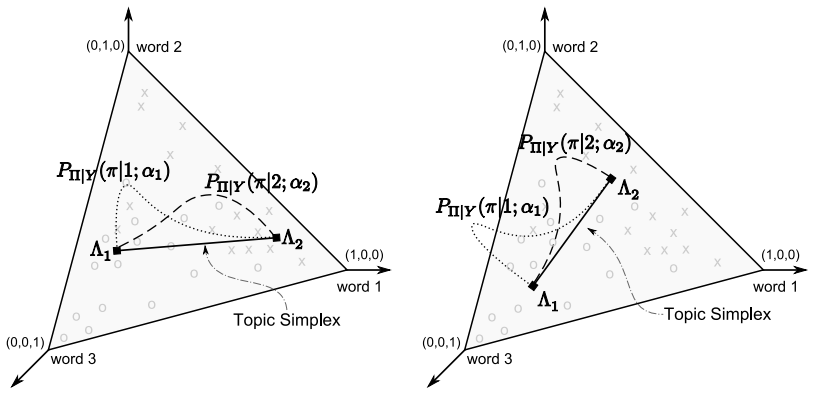
sample a class label $w \sim P_{W|Z}(w|\bar{\mathbf{z}}; \boldsymbol{\zeta}_{1:C})$, $w \in \mathcal{W}$

end for

where, $\bar{\mathbf{z}}$ is the mean topic assignment vector $\bar{\mathbf{z}}_k = \frac{1}{N}\sum_{n=1}^{N} \delta(z_n, k)$, and

$$P_{W|Z}(w|\bar{\mathbf{z}}; \boldsymbol{\zeta}) = \frac{\exp(\boldsymbol{\zeta}_w^T \bar{\mathbf{z}})}{\sum_{l=1}^{C} \exp(\boldsymbol{\zeta}_l^T \bar{\mathbf{z}})} \tag{7.1}$$

a softmax activation function with parameter $\boldsymbol{\zeta}_c \in \mathbb{R}^K$. The parameters of this model can be learned with variational inference, as described in [167].



(a) cLDA (b) ts-cLDA

Figure 7.2: Representation of cLDA and ts-cLDA on a three *word simplex*. Also shown are sample images from two classes: "o" from class-1 and "x" from class-2. a) cLDA model with two topics. The line segment depicts a one-dimensional *topic simplex,* whose vertices are topic-conditional word distributions. Each class defines a smooth distribution on the topic simplex, denoted by the contour lines. c) ts-cLDA model. Topic-conditional word distributions are learned with supervision which encapsulate the class attributes.

7.2.4 Geometric Interpretation

The models discussed above have an elegant geometric interpretation [14, 139]. Given a vocabulary of $|\mathcal{V}|$ distinct words, a $|\mathcal{V}|$ dimensional space can be constructed where each axis represents the occurrence of a particular word. A standard $|\mathcal{V}|-1$-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 7.2(a) illustrates the two dimensional simplex of all probability distributions over three words. Also shown are some sample images from two classes, "o" from class-1 and "x" from class-2.

Figure 7.2(a) 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 add to one, a set of K topics defines a $K-1$ simplex, here denoted the *topic simplex*. When the number of topics K is smaller than the number of words $|\mathcal{V}|$, the topics span a low-dimensional

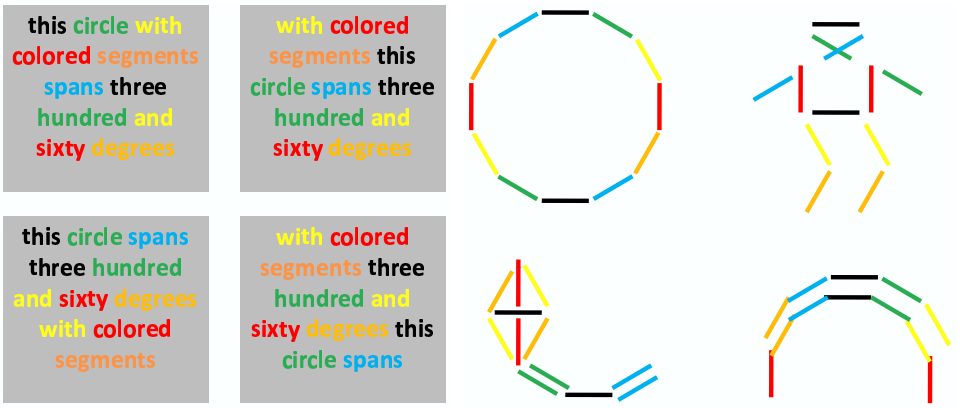


Figure 7.3: left) Four groups of words with equal word histograms. right) Four groups of edge segments with the equal edge segment histograms. Note that each group can be derived from the others by a displacement of words or edge segments. (This figure is best viewed in color)

sub-simplex of the word simplex. The projection of images on the topic simplex can be thought of as *dimensionality reduction*. In Figure 7.2(a), the two topics are represented by Λ_1 and Λ_2, and span a one-dimensional simplex, shown as a connecting line segment. In cLDA, each class defines a distribution (parameterized by $\boldsymbol{\alpha}_w$) 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 also be represented on the topic simplex, where each class defines a softmax function[3].

7.3 The Importance of Supervision

The architecture of holistic context models bear close resemblance to that of cLDA. In Section 2.3, it was shown that SMNs can be computed using the graphical model of Figure 2.5(b). In fact, the graphical model of Figure 2.5(b) *is* that of LDA. Holistic context models introduce a second layer of modeling, using

[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.

multi-modal Dirichlet distribution, on top of the SMNs obtained using the LDA framework. This is similar in principle to the cLDA model where a uni-modal Dirichlet distribution is introduced. Figure 7.1(b) presents the complete version of this model, including the concept variable W at the semantic level. Given the equivalence of the graphical models, it is worth discussing in detail the differences between the two approaches. The fundamental difference is the *level of abstraction* of the intermediate stage of the representation (topics vs. SMNs). While topics are learned in an unsupervised manner, SMN features have *explicit* semantics.

Recall the *semantic gap* between appearance features and visual classes. While text features (words) are intrinsically semantic, this is *not* the case for vision, where localized appearance features (e.g. edge segments) *have no semantic interpretation*. This is illustrated in Figure 7.3, where we present four groups of text (words) and appearance (edge segments) features *with identical distributions*. Because the word features are semantic, it is very difficult to construct a group (sentence) with the same words that is semantically far from the others. This is absolutely not the case for vision, where equivalence of feature distributions places almost no constraint on the group semantics. As the figure shows, the exact same segments can very easily be used to construct groups that depict completely unrelated concepts. The fact that *equivalence of feature distributions does not translate into semantic equivalence* is denoted a semantic gap.

While the semantic gap is small for text (semantic features), it is large for images. Thus, the success of a representation for text classification is an unreliable predictor of its success for scene classification. In particular, the observation that unsupervised topic discovery produces semantic topics for text [14, 58], is very weak evidence that it will be successful for visual recognition. In fact, Figure 7.3 shows that it cannot. In the absence of explicit supervision for topic semantics, it is impossible to learn that the four edge groupings of (c) belong to different topics. On the contrary, the four groups form a perfect appearance cluster, since their segment histograms *are identical*. Unfortunately, due to the semantic gap, this cluster has no well defined semantics *as a whole*. Hence, unsupervised topic learning has no ability to bridge the semantic gap between local appearance and

visual classes. This is unlike the proposed architecture, where SMN features are learned with explicit supervision, and it does make sense to talk about a *semantic space*.

It should be emphasized that in this toy example, although explicit topic supervision results in four classes of *identical* distribution (a highly suboptimal clustering under any unsupervised learning criteria), it produces the *semantically correct* statistical description of the data under the chosen image representation. Note that, under this model, all images of Figure 7.3(right) have an equal chance of being assigned to any of the classes. This is a classifier of higher probability of error than that learned without supervision. In fact, it is the weakest possible classifier. On the other hand, unsupervised topic modeling produces a much stronger classifier: all images assigned to one class with high probability, other classes mostly noise. In summary, the supervised model reflects *both* the true semantics of the data and the ambiguity of the image representation. It attempts to perform the *right* classification but can only do so with high uncertainty. The unsupervised model *invents* an alternative classification problem, which has nothing to do with the image semantics but *can be* solved very accurately. In addition to producing a semantically useless image description, it is also confident on its accuracy.

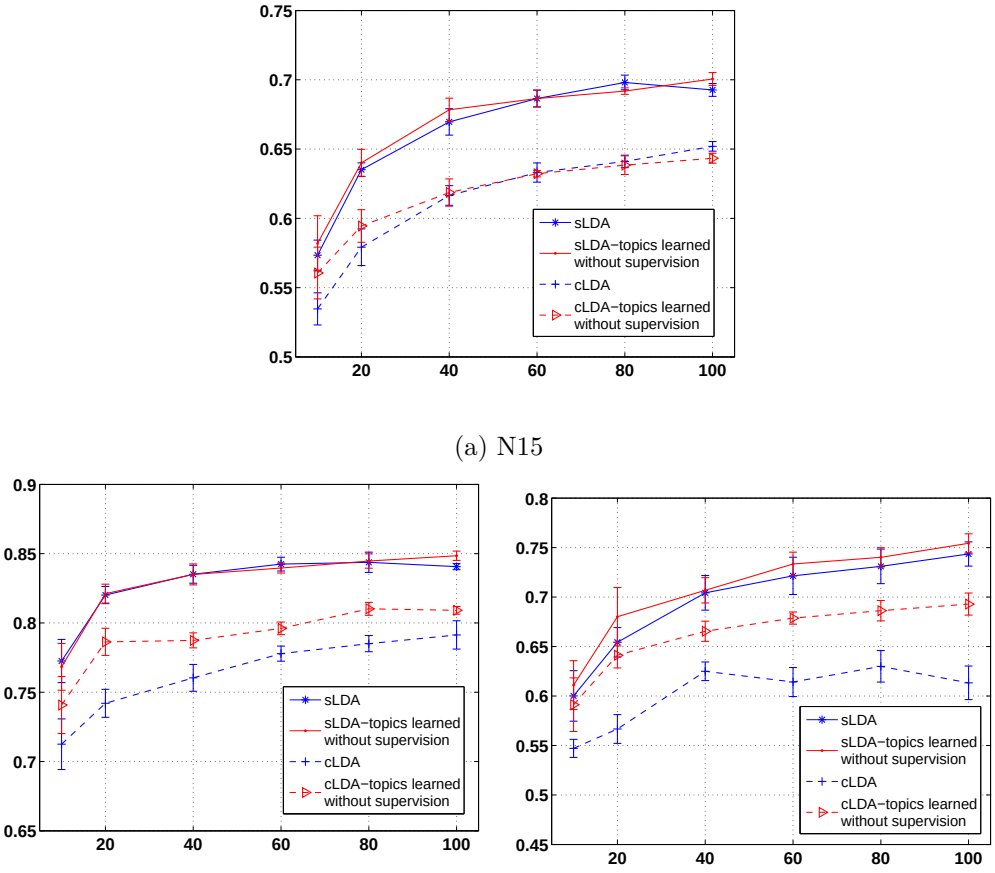
7.4 Limitations of Existing models

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.

7.4.1 Theoretical Analysis

We start by showing that, in both cLDA and sLDA, the class label has a very weak influence in the learning of topic distributions. This is accomplished by an analysis of the learning equations for both cLDA and sLDA, using the variational approximation framework.

In both sLDA and cLDA the parameters $\Lambda_{1:K}$ of the topic distributions are



(a) N15

(b) N8

(c) S8

Figure 7.4: 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 (a) N15, (b) N8 and (c) S8. Similar behavior was observed for codebooks of different sizes.

obtained via the variational M-step as:

$$\Lambda_{kv} \propto \sum_{d} \sum_{n} \delta(v_n^d, v) \phi_{nk}^d \tag{7.2}$$

where d indexes the images, $\sum_v \Lambda_{kv} = 1$, $\delta()$ is a Kronecker delta function and ϕ_{nk} is the parameter of the variational distribution $q(z)$. This parameter is computed

in the E-step with

$$\text{For cLDA:} \qquad \gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_{w^d k} \tag{7.3}$$

$$\phi_{nk}^{d*} \propto \Lambda_{kv_n^d} \ \exp\left[\psi(\gamma_k^d)\right] \tag{7.4}$$

$$\text{For sLDA:} \qquad \gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_k \tag{7.5}$$

$$\phi_{nk}^{d*} \propto \Lambda_{kv_n^d} \exp\Bigg[\psi(\gamma_k^d) + \frac{\zeta_{w^d k}}{N} - \frac{\sum_c \exp\frac{\zeta_{ck}}{N} \prod_{m\neq n} \sum_j \phi_{mj}^d \exp\frac{\zeta_{cj}}{N}}{\sum_c \prod_m \sum_j \phi_{mj}^d \exp\frac{\zeta_{cj}}{N}}\Bigg] \tag{7.6}$$

where, $\boldsymbol{\gamma}$ is the parameter of the variational distribution $q(\boldsymbol{\pi})$ (see [14] for the details of variational inference in LDA). The important point to note is that the class label w^d only influences the topic distributions through (7.3) for cLDA (where $\boldsymbol{\alpha}_{w^d}$ is used to compute the parameter $\boldsymbol{\gamma}^d$) and (7.6) for sLDA (where the variational parameter ϕ_{nk}^d depends on the class label w^d through $\zeta_{w^d k}/N$).

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 γ_k has two components: $\hat{l}_k = \sum_n \phi_{nk}$, which acts as a vector of counts, and α_{wk} which is the parameter from the prior distribution. As the number of samples increases, the amplitude of the count vector, $\hat{\mathbf{l}}$, increases proportionally, while the prior $\boldsymbol{\alpha}_w$ remains constant. Hence, for a sufficiently large sample size N, the prior $\boldsymbol{\alpha}_w$ 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 W and the learned topics Z_i is *extremely weak*. This is not a fallacy of the variational approximation. In cLDA (Figure 7.1(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 $\zeta_{w^d k}/N$. This clearly diminishes as the sample size N increases. 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.

7.4.2 Experimental Analysis

To confirm the observations above, we performed experiments with topics learned under two approaches. In the first, we used the original learning equations, i.e. (7.3) and (7.4) for cLDA and (7.5) and (7.6) for sLDA. In the second we severed all connections with the class label variable *during learning* (of the topics), by reducing the variational E-step (for both cLDA and sLDA) to,

$$\gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha \tag{7.7}$$

$$\phi_{nk}^{d*} \propto \Lambda_{kv_n^d} \exp\left[\psi(\gamma_k^d)\right] \tag{7.8}$$

with $\alpha = 1$. This guarantees that the topic-conditional distributions are learned without any class influence. The remaining parameters ($\boldsymbol{\alpha}_w$ for cLDA, $\boldsymbol{\zeta}_w$ 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 7.4 shows the scene classification performance of cLDA and sLDA, under the two learning approaches, on the N15, N8, and S8 datasets (see Appendix A for details on the experimental setup). 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, on the three datasets. For cLDA, unsupervised topic discovery is in fact *superior* on the N8 and S8 dataset. This can be explained by poor regularization of the original cLDA algorithm. We have observed small values of α_{wk}, which probably led to poor estimates of the topic distributions in (7.3). For example, the maximum, median and minimum values of α_{wk} learned with 10 topics on S8 were 0.61, 0.12, 0.04 respectively. In contrast, the corresponding values for

unsupervised topic discovery were 7.09, 1.09, 0.55. Similar effects were observed in experiments with codebooks of different size. These results are clear evidence that the performance of cLDA and sLDA is similar (if not inferior) to that of topic learning without class supervision. In both cases, the class variable has very weak impact on the learning of topic distributions.

7.5 Topic supervision

In this section introduce topic supervision for LDA models, and its impact in learning and inference.

7.5.1 Topics supervision in LDA model

The simplest solution to the limitations discussed in the last section, is to *force* topics to reflect the semantic regularities of interest. This consists of equating topics to class labels, and is denoted *topic supervised LDA*. Topic supervision was previously proposed in semi-LDA [170] and labeled-LDA [116], for action and text classification respectively. However, its impact on classification performance is difficult to ascertain from these works, for several reasons. First, none of them performed a systematic comparison to existing LDA methods. Second, both are topic-supervised versions of LDA. Intuitively, topic supervised versions of classification models, namely cLDA and sLDA, should achieve better performance. Third, semi-LDA adopts an unconventional inference process, which assumes that $p(z_n|v_1, v_2, \ldots, v_n) \propto p(z_n|\boldsymbol{\pi})p(z_n|v_n)$. It is unclear how this affects the performance of the topic-supervised model. Finally, the goal of labeled-LDA is to assign multiple labels per document. This is somewhat different from scene classification, although labeled-LDA reduces to a topic-supervised model for classification if there is a single label per item.

7.5.2 Models and geometric interpretation

To analyze the impact of topic-supervision on the various LDA models, we start by noting that 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*. It is thus possible to introduce topic-supervised versions of all models in the literature. In this work, we consider three such versions, viz. "topic supervised LDA (ts-LDA)", "topic-supervised class LDA (ts-cLDA)", and "topic-supervised supervised LDA (ts-sLDA)". These are the topic-supervised versions of LDA, cLDA and sLDA, respectively, with the following three distinguishing properties,

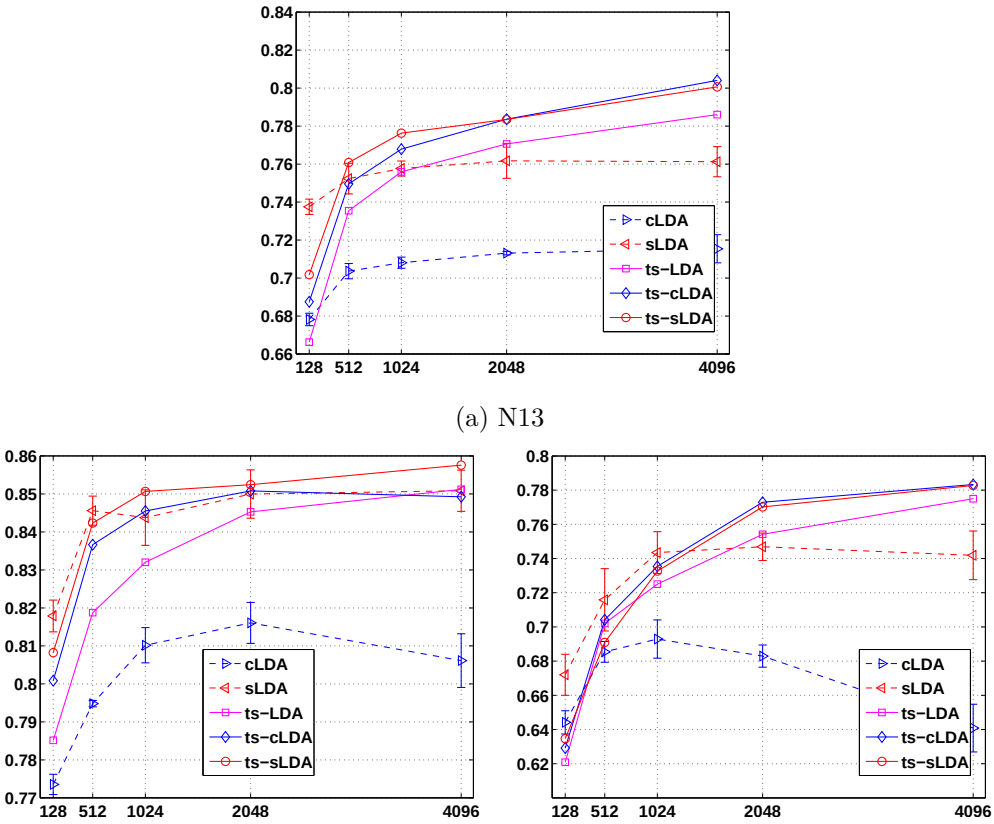
- the set of topics $\mathcal{L}$ *is* the set of class labels $\mathcal{W}$.
- the samples from the topic variables Z_i are class labels.
- the topic conditional distributions $P_{V|Z}()$ are learned in a supervised manner.

We will shortly see that this has the added advantage of substantially simpler learning.

Figure 7.2(b) shows the schematic of ts-cLDA for a two class problem on a three word simplex. As with cLDA, Figure 7.2(a), Λ_1 and Λ_2 represent two topic-distributions. There is, however, a significant difference. For cLDA, topic distributions are learned in a bottom up manner and can be positioned anywhere on the word simplex, by the topic discovery algorithm. For ts-cLDA, the topics are specified: each topic is an image class.

7.5.3 Learning and inference with topic-supervision

The introduction of topic-level supervision decouples the learning of the topic-conditional distribution $P_{V|Z}()$ from that of the other model parameters, substantially reducing learning complexity. In general, learning topic distributions would require a strongly supervised training set, however in absence of these labels, all patch labels in an image are made equal to its class label, i.e. $z_n^d = w^d \ \forall n, d$.



(a) N13

(b) N8

(c) S8

Figure 7.5: Performance of ts-sLDA, ts-cLDA, sLDA, and cLDA as a function of codebook size on (a) N13, (b) N8 and (c) S8. 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.

This type of learning has shown to be effective, both through the design of image labeling systems [21] and theoretical connections to multiple instance learning [155]. The ML estimate of Λ_k is

$$\Lambda_{kv}^{*} = \arg\max_{\Lambda_k} \sum_{d} \sum_{n} \delta(w^d, k)\delta(v_n^d, v) \log \Lambda_{kv} \tag{7.9}$$

such that $\sum_{v=1}^{|\mathcal{V}|} \Lambda_{kv} = 1$. The solution to this optimization problem is

$$\Lambda_{kv} = \frac{\sum_d \sum_n \delta(w^d, k)\delta(v_n^d, v)}{\sum_j \sum_d \sum_n \delta(w^d, j)\delta(v_n^d, v)}. \tag{7.10}$$

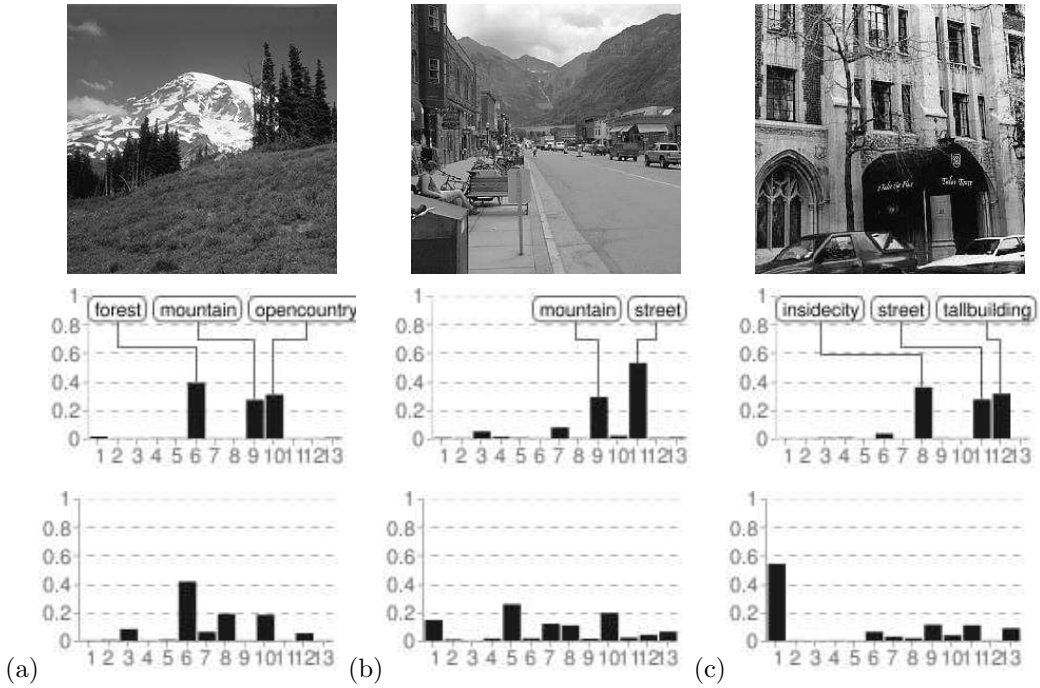


Figure 7.6: Some example images that were misclassified by cLDA, but correctly classified using ts-cLDA. The expected topic distributions for ts-cLDA and cLDA (using 13 topics) are shown in the middle and bottom rows respectively. For ts-cLDA, topic labels are same as the class labels and the high probability topics are indeed the ones which capture the semantic meaning of the image. For cLDA, the topic labels do not carry any clear semantic meaning.

Given the topic-conditional distributions, all other parameters can be learned as in the original models. Parameter estimation for ts-cLDA is detailed in Appendix F.

7.5.4 Experimental analysis

Figure 7.5 presents classification results of ts-LDA, ts-cLDA and ts-sLDA, as a function of codebook size, under the experimental conditions of Figure 7.4. Compared to sLDA and cLDA, all three topic supervised approaches achieve superior classification performance. This is true for all datasets across different codebook size when compared to cLDA, and for all datasets and codebooks with over 1024 codewords when compared to sLDA. The best performance across dif-

Table 7.1: Classification Results on Natural Scene Categories.

	Dataset		
model	**N15**	**N13**	**N8**
ts-sLDA	**74.82 ± 0.68**	**79.70 ± 0.48**	**86.33 ± 0.69**
ts-cLDA	74.38 ± 0.78	78.92 ± 0.68	86.25 ± 1.23
ts-LDA	72.60 ± 0.51	78.10 ± 0.31	85.53 ± 0.41
sLDA	70.87 ± 0.48	76.17 ± 0.92	84.95 ± 0.51
cLDA	65.50 ± 0.32	72.02 ± 0.58	81.30 ± 0.55

ferent codebooks and topics cardinality is reported in 7.1 and 7.2. On average, across datasets, topic-supervision improves the classification accuracies of cLDA and sLDA by 12% and 5% respectively. Among the three topic-supervised models, ts-cLDA and ts-sLDA achieve comparable performance, which is superior to that of the simpler ts-LDA model.

Figure 7.6 shows some images incorrectly classified by cLDA but correctly classified by ts-cLDA, on the N15 dataset. Also shown are the topic histograms obtained in each case, with ts-cLDA in the middle and cLDA in the bottom row. The figures illustrate the effectiveness of ts-sLDA at capturing semantic regularities — topics with high probability are indeed representative of the image semantics. Note that such an interpretation is only possible as the topic labels in ts-cLDA have a one-to-one correspondence with the class labels. For cLDA, topic histograms merely represent visual clusters.

7.6 Acknowledgments

The text of Chapter 7, in full, is based on the material as it appears in: N. Rasiwasia and N. Vasconcelos, *'Holistic Context Models for Visual Recognition'*, Accepted to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence, and N. Rasiwasia and N. Vasconcelos, *'Generative Models for Image Classification'*, In preparation for IEEE Transactions on Pattern Analysis and Machine Intelligence. The dissertation author was a primary researcher and an author

Table 7.2: Classification Results on Sports8 and Corel50.

	Dataset	
model	**S8**	**C50**
ts-sLDA	**78.37 ± 0.80**	**42.33**
ts-cLDA	77.43 ± 0.97	40.80
ts-LDA	77.77 ± 1.02	39.20
sLDA	74.95 ± 1.03	39.22
cLDA	70.33 ± 0.86	34.33

of the cited material.